CAPACITY PLANNING AND SCHEDULING WITH APPLICATIONS IN HEALTHCARE

A Thesis
Presented to
The Academic Faculty

by

Monica Cecilia Villarreal

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
May 2015

CAPACITY PLANNING AND SCHEDULING WITH APPLICATIONS IN HEALTHCARE

Approved by:

Professor Pinar Keskinocak, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Turgay Ayer
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor David Goldsman
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Paul Griffin
Harold and Inge Marcus Department of Industrial and Manufacturing Engineering
Pennsylvania State University

Professor Julie Swann
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: 4 December 2014

To my parents,

for their unconditional love, faith, and support.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Pinar Keskinocak for her invaluable guidance during the doctorate program, not only in academic, but also in professional and personal matters. Her passion, optimism, and confidence in everything she does is really inspiring and has motivated me to be a better student, researcher and professional. I especially would like to thank Dr. Keskinocak for her continuous support, which gave me strength through the most challenging times of this program. I also would like to thank the rest of my thesis committee members, Dr. Julie Swann for her support and advice, and for being there for me when I needed it; Dr. David Goldsman for his valuable insights (particularly during the development of Chapter 3), and for his contagious postive energy and humor, in and out the classroom; and Dr. Turgay Ayer and Dr. Paul Griffin for their helpful comments about this work. I am specially grateful to Stanley Bartlett, Marilyn Bowcutt, Lisa Jackson, Christine Martin, Susan McMillan and Jonathan Turner from University Hospital in Augusta, Georgia, for their support and involvement in the research presented in Chapters 2 and 4.

I want to thank the faculty, administrators, and students from the H. Milton Stewart School of Industrial and Systems Engineering (ISyE) for their support and friendship. In particular, I would like to thank Dr. Jim Dai for giving me a confidence boost and the reassurance that with hard work I could achieve my goals; David Cowan from the Health Systems Institute, for the inspiring discussions about today's challenges of the healthcare system; fellow ISyE students Dr. Camilo Ortiz, Dr. Todd Levin, Dr. Vinod Cheriyan, Dr. Pengyi Shi, Dr. Antonio Carbajal, Dr. Jon Petersen, Dr. Tuba Yilmaz and Alvaro Lorca, for their help and feedback during the program;

and my office neighbors Tugce Isik and Dr. Mallory Soldner for being there when I needed a break after long days of work.

I also want to thank my dear Atlanta friends who have always been there for me during this process, including Tatiana Restrepo, Carlos Campos, Kay Perez, Alejandro Suarez, Barbara Pina, Ciro Espinoza, Felipe Castrillon, David Gonzalez, Giovanni Rosso, Lubna Rashid, Robby and Brittany Espinosa, Tito and Pilar del Valle, Gina Lee, among many others. Especially, I want to thank my roommates Dr. Maria Restrepo and Vivian Soler, for being like family to me. My infinite gratitude goes to my dear husband, who has been my rock. In addition, I want to thank my husband's family, in particular Paul Cella and family, for all the great times and dinners we enjoyed together.

I am forever indebted to my parents and family, for giving me their unconditional support and love since I can remember, no matter the distance between us. I would not be where I am now without them. Last, but not least, I want to thank God who has graced my life with great people and opportunities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

In this thesis we address capacity planning problems with different demand and service characteristics, motivated by healthcare applications. In the first application, we develop, implement, and assess the impact of analytical models, accompanied by a decision-support tool, for operating room (OR) staff planning decisions with different service lines. First, we propose a methodology to forecast the staff demand by service line. We use these results in a two-phase mathematical model that defines the staffing budget for each service line, and then decides how many staff to assign to each potential shift and day pair while considering staff overtime and pooling policies and other staff planning constraints. We also propose a heuristic to solve the model's second phase. We implement these models using historical data from a community hospital and analyze the effect of different model parameters and settings. Compared with the current practice, we reduce delays and staff pooling at no additional cost. We validate these conclusions through a simulation model.

In the second application, we consider the problem of staff planning and scheduling when there is an accepted time window between each order's arrival and fulfillment, with the goal of obtaining a balanced schedule that focuses on on-time demand fulfillment but also considers staff characteristics and operational practices. Hence, solving this problem requires simultaneously scheduling the staff and the forecasted demand. We propose, implement, and analyze the results of a model for staff and demand scheduling under this setting, accompanied by a decision-support tool. We implement this model in a company that offers document processing and other back-office services to healthcare providers. We provide details on the model validation, implementation, and results, including a 25% increase in the company's staff productivity.

Finally, we provide insights on the effects of some of the model's parameters and settings, and assess the performance of a proposed heuristic to solve this problem.

In the third application, we consider a non-consumable resource planning problem. Demand consists of a set of jobs, each job has a scheduled start time and duration, and belongs to a particular demand class that requires a subset of resources. Jobs can be 'accepted' or 'rejected,' and the service level is measured by the (weighted) percentage of accepted jobs. The goal is to find the capacity level that minimizes the total cost of the resources, subject to global and demand-class-based service level constraints. We first analyze the complexity of this problem and several of its special cases, and then we propose a model to find the optimal inventory for each type of resource. We show the convergence of the sample average approximation method to solve a stochastic extension of the model. This problem is motivated by the inventory planning decisions for surgical instruments for ORs. We study the effects of different model parameters and settings on the cost and service levels, based on surgical data from a community hospital.

CHAPTER I

INTRODUCTION

The healthcare expenditures in the United States reached $3.8 trillion dollars in 2013 [97], and are expected to increase at a pace faster than the economic growth. Therefore, there is a strong incentive to control healthcare costs [100]. This represents a great opportunity for the application of operations research and management science (OR/MS) methods and tools to improve the efficacy and efficiency of the healthcare system through better decision making [102, 111]. One of the questions OR/MS can help to answer is: how can we make better use of the limited (and often expensive) resources in the healthcare system? These resources include staff (clinical and administrative), space (hospitals, clinics), equipment and instrumentation, etc. In this thesis, we focus on three applications related to resource capacity planning and management with applications in healthcare. Each chapter of this thesis covers one of these applications, and all chapters can be read independently. We address both strategic (i.e, resource forecasting and planning) and tactical (i.e, planning the execution of the delivery process: who, what, how, where and when) aspects of capacity planning and resource allocation problems in healthcare [74].

The first two applications are related to staff planning and scheduling. Labor is a major cost component in many industries. Salaries are commonly the main component of all the operating expenses in service industries such as healthcare (about 52%) [110]. For this reason, workforce planning and scheduling is crucial, since an improvement in labor productivity and staff satisfaction can translate into significant savings and reduce staff turnover. There is an abundant literature on staff planning

and scheduling. Ernst et al. [58] give an annotated bibliography of about 700 references ranging from 1954 to 2004. Our first application is related to the staff planning for a surgical department. The surgical department is one of the most expensive departments in a hospital, consuming about 10% of the hospital's budget [67, 117], and it has significant impact in the hospital's overall operations as 60% of the inpatient admissions result from a visit to the operating room (OR) [115]. The ORs work under a surgical schedule and insufficient staff can lead to delays in scheduled surgeries, which can result in dissatisfaction among both patients and surgeons, and an increase of overtime. On the other hand, limiting the number of staff to achieve high staff utilization is also important. Hence, OR staff planning decisions need to balance difference objectives (including minimizing delays and maximizing staff utilization), constraints, and preferences. In Chapter 2, we present a two-phase model for a staff planning problem in a surgical department. We consider the setting where staff, in particular nurse circulators and surgical scrub technicians, are assigned to a service line, and while they can be 'pooled' and temporally assigned to another service line if needed, these re-assignments should be limited. In Phase I, we decide on the number of staff hours to budget for each service line, considering policies limiting staff pooling and overtime, and different demand scenarios. In Phase II, we determine how these budgeted staff hours should be allocated across potential work days and shifts, given estimated staff requirements and shift-related scheduling restrictions. We propose a heuristic to speed the model's Phase II solution time. We test these models using historical data from a 580-bed, not-for-profit, community hospital, which performs about 8,000 surgical procedures annually in 14 ORs, and compare the model's results with the hospital's current practices. Using a simulation model for the surgical operations, we find that our two-phase model reduces both staff pooling and the delays caused by staff unavailability, without increasing the workforce size. Finally, we describe a decision-support tool we developed with the objective of fine-tuning staff

planning decisions.

In the second application, we change our focus to healthcare back office staff. This research is motivated by the staff planning and scheduling decisions of a company that offers document processing and other back office services to healthcare providers. Back office operations are rarely discussed in the OR/MS literature, but healthcare providers spend over $100 billion to manage claims as more than half of healthcare transactions are still paper-based and manually processed [14]. In addition, efficiency and efficacy in back office healthcare operations are crucial as bad debt expenses average 12% to 13% of revenue, and the bills collection cycle averages more than 90 days [94]. The company that motivated our study requires a staff schedule which focuses on on-time demand fulfillment but also considers staff preferences and operational practices. In Chapter 3, we develop a mathematical model for planning and scheduling staff and demand considering a time window for on-time demand fulfillment, and staff individual characteristics, preferences, and availability. We also discuss a version where the staff schedule is fixed. The model can be applied in many service settings such as general back office services, warehouses, and fulfillment centers. We develop a user-friendly decision-support tool that employs the model and the solution methodology, and implement it in the healthcare back-office services provider, considering additional operational practices of this company such as team leaders scheduling. We conduct a computational study and develop insights regarding the trade-offs between the on-time demand fulfillment and the quality of the staff schedule, the effect of a change in the time window, the impact of client behavior (e.g., batch arrivals of demand), and the consequences of considering additional preferences and operational constraints in the model. We also evaluate the robustness of the staff schedule generated by the model under different demand scenarios. Finally, we present a heuristic to find high quality staff schedules quickly. After the implementation, the company reported a 25% increase in staff productivity.

In Chapter 4, we introduce a resource planning problem that arises in systems with non-consumable resources. Demand consists of a set of jobs, where each job has a scheduled start time and a duration. There are multiple job types, each corresponding to a particular demand class and requiring a predefined subset of resources to be completed. Jobs can be 'accepted' or 'rejected'. The goal is to minimize the cost of resources (i.e, deciding on the level of inventory/capacity for resources), subject to constraints on service levels, measured by the (weighted) percentage of accepted jobs (globally and per class). This problem is motivated by hospital operations, in particular, by the instruments planning of a surgical department, where most of the surgical cases (jobs) are scheduled in advance according to surgeons' and patients' preferences and staffed ORs availability. The average U.S. hospital has a surgical instruments inventory worth approximately between $2 and $4 million dollars, and there are a lot of opportunities for improvement and cost reduction though better planning [64]. Similar problems also arise in some applications of workforce planning, and repair and maintenance operations. We first present complexity results for various special cases of this problem, and then propose a model to find the optimal capacity for each type of resource with service constraints considered at the global and demand class levels. We also propose a stochastic extension of the model for the case where the capacity/resource decisions need to be done before the demand schedule is revealed. In this case, the resource capacities should meet expected service levels. We propose a Sample Average Approximation (SAA) approach for this stochastic program and show its convergence. Finally, with the goal of studying the effects of different model parameters and settings on cost and service levels, we develop a case study based on surgical data from the community hospital introduced in Chapter 2.

CHAPTER II

STAFF PLANNING FOR OPERATING ROOMS WITH DIFFERENT SURGICAL SERVICES LINES

2.1 Introduction

The surgical department operations have a major impact across a hospital as about 60% of the inpatient admissions result from a visit to the operating room (OR) [115]. Moreover, the surgical department is one of the most expensive departments in a hospital, consuming about 10% of the hospital's budget [67, 117]. An adequate OR staff planning that balances different objectives and costs is important. Given that the ORs work under a schedule of surgical procedures, insufficient staff can lead to procedure start time delays, which can result in dissatisfaction among patients and surgeons, as well as an increase of overtime. On the other hand, a staff surplus can result in lower staff utilization. The aim of this chapter is to develop, implement, and assess the impact of analytical models for OR staff planning decisions. We test these models using historical data from a 580-bed, not-for-profit, community hospital, which performs about 8,000 surgical procedures annually in 14 ORs.

This surgery department uses both open access and *block schedule*. A time block in the OR schedule is a pre-allocation of OR time to a particular *surgical service*, that could be offered to other services at some point based on availability and proximity to the day of surgery. A surgical service is composed of a surgeon or set of surgeons who perform surgical procedures related to a particular set of surgical specialties. Surgeons have for the most part their own practices and are not directly employed by the hospital but rather schedule their surgeries during the offered OR times. On the other hand, anesthesiologists, patient care assistants, nurse circulators, and surgical

scrub technicians (scrub techs) are usually employed by the hospital and remain in the hospital (or on-call) during the length of their shifts. In this chapter we focus on the staffing decisions regarding circulators and scrub techs (hereafter referred to as *OR staff* or *staff*), which account for about a third of the OR time costs [49].

In 2011, the surgical department of this hospital formally established three main groups or lines of surgical services or *service lines*: cardio/vascular, neurology/orthopedic (neuro/ortho), and general surgery (general). Aligned to these service lines, the surgical department also established a new staffing strategy such that OR staff should work exclusively or mostly on cases of their assigned service line. The motivation behind this initiative was that if surgeons consistently work with the same staff teams, their satisfaction would increase as they would become more comfortable with their teams, and as this *staff specialization* by service line would help OR staff to improve their expertise within a particular set of surgical services. OR staff could still be assigned to the cases of other service lines; however, such assignments (*staff pooling*) should be limited or avoided if possible. For this reason, the hospital planned to hire new staff, including an OR coordinator by service line, who would be responsible for the OR staff scheduling and case assignment for the service line.

Staff specialization is motivated by the fact that a particular set of skills may be necessary to cover one or a subset of different types of demand. However, a pure staff specialization strategy requires a match of supply and demand within each service line, and a mismatch may result in delays, loss of demand, or idle time. A staff pooling strategy allows staff to cover different types of demand to create a better match between demand and supply. This would be particularly effective if the overall demand is relatively stable and the demand variability occurs mainly within a service line. The hospital's new service-line-oriented strategy brought additional complexities to the staff planning process given the possibility of sharing staff among service lines, particularly in emergency and add-on cases.

The pressure of keeping current and attracting new surgeons (and their patients) on one side, and keeping control of the operating costs on the other, brings out the need for an effective staff planning methodology that minimizes costs while achieving a desirable performance. In this chapter, we propose a methodology to forecast weekly demand requirements and to estimate the required staff at any time of the week, for each surgical service line. Then, we use these results in a two-phase model to support the decision making process of the OR manager: define the number of full-time equivalent employees (FTEs) required by service line (i.e., the *staffing budget*), followed by defining an adequate *staffing structure*: how many staff to assign to each potential shift and day pair and service line (e.g., two cardio/vascular scrub techs on Mondays from 7:00AM to 3:00PM) while considering staff overtime and pooling policies, and other staffing constraints. We analyze the benefit of implementing this two-phase model versus the hospital's current practices and provide managerial insights regarding the impact of overtime and pooling policies. We also analyze the effect of including additional constraints such as limiting the number of different shifts or penalizing deviations from the current staffing structure.

In Section 2.2 we discuss some relevant literature. In Section 2.3 we describe the hospital's current staff planning process and introduce a two-phase staff planning model. In Section 2.4 we discuss the model implementation and the use of surgical data to estimate the staff demand input data by service line. We also present the results of such implementation, describe and evaluate a heuristic for the model's second phase, and present a simulation approach to evaluate the results. Finally, in Section 2.5 we present conclusions of this research and propose potential directions.

2.2 Literature Review

Most of the work on staff planning and scheduling in healthcare settings focuses on nurse scheduling. Bard [16] conducts a review of nurse scheduling models in the literature, and suggests that the staff planning and scheduling problems can be decomposed hierarchically in: (1) long-term planning (fix the composition of the permanent workforce), (2) medium-term scheduling (determine shifts and days-off assignments), where the majority of the literature has focused on; (3) short-term scheduling (assign tasks to workforce, manage vacations, overtime, casuals, part-time workforce), and (4) real-time control (adjust workforce for disruptions such as emergencies and absenteeism). Bard concludes that there is a need of robust treatment of demand at each of the hierarchical levels as well as a better understanding of the long-term staffing requirements. This research answers to both needs: It proposes and evaluates a methodology that starts by analyzing historical demand data from a surgical department and then uses these results as input for a long-term staff planning model.

There is extensive literature on staff planning and scheduling with several applications in and outside of healthcare (see [5, 114, 58] for reviews). Ernst et al. [57] reviewed over 700 papers in personnel scheduling, from which about 60% focus on crew or tour scheduling, whereas less than 15% focus on staff planning (long-term). Among the earlier papers on staff planning are [11] and [12]. The authors develop exact expressions to compute minimum workforce bounds for the days-off scheduling problem, considering a variable daily demand and a single shift. These bounds and those in [119] are revisited by Alfares et al. [4] and incorporated in a mathematical program for the five workdays and two days-off scheduling problem. Mathematical programs have been widely used to solve more general staff planning and tour scheduling problems where demand varies during a shift and shifts can start at different times [21, 28, 35]. Because the number of potential *tours* (a sequence of shifts that can be assigned to a worker) grows fast with the number of different shift start times and

durations, Bard et al. [17] propose to first solve the staffing structure for the planning horizon and then post-process the results to obtain tours. In this research, we focus on planning for this staffing structure. Sinreich et al. [108] propose considering shifts with many potential start times and durations, i.e., staggering work shifts, as a way of downsizing the workforce in emergency rooms; however, they do not consider different service lines, constraints on the minimum number of staff required to cover these shifts, or variable demand patterns during the planning horizon. Demand uncertainty has been incorporated by means of stochastic programming [18, 65]. Queueing and simulation models have been used to compute the staff demand at different times, which we refer to as *staffing level* requirements [2, 38, 121].

Staff pooling strategies and staff cross-training have been proposed in the literature. Camm et al. [31] propose a central staff pool from which additional staff can be pulled from when assigned staff are not sufficient to cover a certain type of demand. Bard et al. [15, 19] and Brunner et al. [27] propose that more skilled staff can be used to substitute for less skilled staff, also known as downgrading. Additionally, cross-trained staff can be reassigned to cover all or a subset of demand types [37, 56, 77, 105]. In this research, we focus on the latter since it reflects more closely the hospital's situation. Li et al. [86] consider cross-training when defining the workforce size for a clinic; however, they do not consider demand uncertainty or variability. Maenhout et al. [89] also consider cross-training and long-term staffing decisions such as staff allocation across different wards in their nurse scheduling model, but the workforce size is given and demand uncertainly is not considered.

A comprehensive literature review on OR planning and scheduling was done by Cardoen et al. [32]. Previous research mostly focuses on scheduling the ORs and making decisions on the day of surgery, such as re-sequencing or assigning cases to ORs [48, 67], and literature on OR staffing decisions is very limited. News vendor approaches that balance under- and over-utilization costs have been proposed for

service-specific OR planning and scheduling to determine the hours into which cases are scheduled and then to staff accordingly [44, 51, 91]. More closely related to our problem setting is the literature on OR staffing during nights, weekends, and holidays, where pooled surgical services analyses are common. Dexter et al. [53, 46] propose to define the set of potential staffing solutions for the weekends and holidays, and then compare the resulting staffing levels with those required given historical data. The solutions with inadequate staffing (when the number of days with case delays is greater than a cut-off level) are discarded, and the remaining potential solutions are evaluated in terms of cost. Dexter et al. [45] use a similar procedure to evaluate the cost impact of increasing the number of OR staff teams during the ORs second shift. van Ostrum et al. [116] use a simulation approach to evaluate potential staffing solutions for the ORs night shift, considering safety intervals for waiting. They model different surgical services, but the OR teams are not explicitly assigned to a particular service or group of services. These staff planning approaches suggest evaluating all potential staffing solutions under historical or randomly generated surgical data, discard those that are not acceptable (e.g., in our case study if the staff overtime or pooling is above the policy limits), and then choose the solution that minimizes the cost. However, the number of potential staffing solutions to test grows exponentially as we consider more potential shifts (with several start times and lengths), all days of the week, different staff types and case staff requirements (rather than a standard OR team for all cases), and the need to assign staff to a particular service line to control for staff pooling. Rather than performing a brute-force or exhaustive search, we propose the use of mathematical programming models to find an optimal OR staffing solution.

Literature in long-term staff planning is limited. To the best of our knowledge this is the only long-term staffing planning model that specifically addresses the concept of different service lines and staff pooling under variable and uncertain demand. In addition, the model consider restrictions on the staffing structure to facilitate its

implementation when scheduling staff. The case study where we apply the model also has some novel features. The staff demand in the ORs comes mostly from scheduled surgeries (except from emergencies and add-on cases), different from other systems with unscheduled demand (emergency rooms, call centers, etc.), for which queueing models for staff planning have been used [76]. This staff demand in the ORs is determined mostly by the OR scheduling practices, such as the use of a block schedule. Many of the challenges in OR staff planning come from the uncertainty and variability of the surgical cases duration and mix, which affect the use of this pre-allocated time. In this research, we describe how to use surgical data to take this into account in the proposed two-phase model for OR staff planning decisions.

2.3 A Two-Phase ORs Staff Planning Model

The staff planning process in the ORs starts with the staffing budget (Phase I), usually aligned with the hospital's budget planning, commonly done every year. The OR manager decides how many permanent FTEs to budget with the goal of meeting the expected demand. An FTE is a measure of labor to make workloads comparable, and it represents the number of hours a standard full-time employee is expected to work per week. To overcome demand fluctuations, overtime could be used. However, overtime is more costly than regular time, there is a limit on how much overtime staff can do, and it can also affect staff satisfaction. Hospitals could also use agency staff to fill in a shift when demand is higher than planned, but this is also expensive and sometimes limited by labor contracts. Phase II decisions focus on determining the staffing structure, i.e., the number of staff assigned to each shift, day, and service line, considering the FTEs available and the demand patterns. The staffing structure should be revised if, for example, scheduling practices, such as the OR block schedule, change and affect the staff requirements, or if the workforce changes. In this setting, staffing structure could be revised every couple months or less frequently.

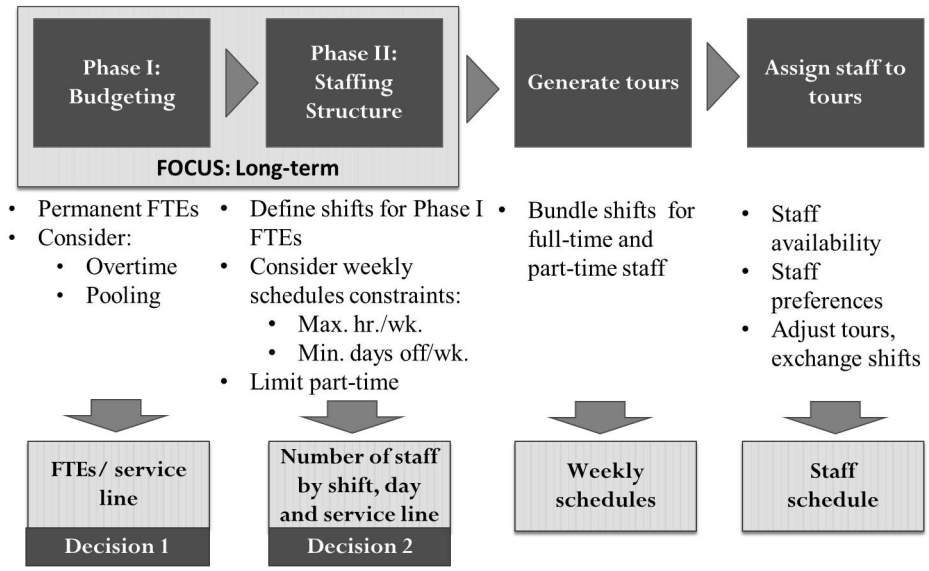


Figure 1: OR staff planning and scheduling process overview: long-term and medium-term decisions.

Following Bard's [16] classification, Phases I and II correspond to long-term staffing decisions. After the staffing structure is specified, the OR coordinators would build valid tours that cover the predefined shifts and assign them to the staff, considering staff availability, preferences, and any scheduling rules (medium-term decisions, outside the scope of this research). Figure 1 shows an overview of both long- and medium-term staff planning and scheduling decisions, including the main activities and outputs at each step of the process, and it highlights the staff planning decisions that are the focus of this research. Finally, each day the OR coordinators assign their staff to a specific OR or to a set of cases according to the OR and staff schedules (short-term decisions), and make adjustments in real time as required (real-time decisions).

To develop a model for the OR long-term staffing decisions and compare its performance to the current practice (CP), we first analyze the hospital's current approach to Phases I and II. The hospital's approach to Phase I is:

1. Establish the baseline data for the next budget cycle (e.g., staffing needs based

on last year's requirements).

2. Estimate the number of cases for each service line considering this baseline.

3. Estimate the expected difference in the number of cases (e.g., for retiring or incoming surgeons) and apply this difference to the baseline.

4. Estimate the required staff hours for each staff type (circulators, scrub techs), based on the *staff productivity* (number of staff hours/case).

5. Compute the number of permanent FTEs based on the required staff hours (2,080 staff hours/FTE).

6. Add relief FTEs to cover vacations and absenteeism (an additional 12%).

The current approach has key limitations: (i) It considers only aggregated baseline data for staff demand and does not incorporate historical trends; (ii) it does not take into account changes in the case mix or in scheduling practices to estimate the required staff hours, even though these changes could affect the staff productivity as staff requirements and case duration can change from case to case; (iii) it does not consider policies regarding overtime and staff pooling, the latter becoming an issue particularly after the establishment of the service-line-based staffing strategy.

For Phase II, the hospital's approach is as follows:

1. Define the OR schedule based on scheduling practices (e.g., considering the daily OR block schedule).

2. Define shifts based on the OR schedule (e.g., OR open from 7:30 AM to 5:30 PM).

3. Assign shifts to a service line, matching the service line with the type of cases performed during the particular OR schedule.

4. Compute staff requirements for each shift, based on the type of cases commonly performed; for instance, cardio and ortho cases require two scrub techs whereas other cases require one.

The first issue with this approach is that the OR schedule is constructed based on the scheduled hours rather than the real hours when the ORs are 'open' for procedures. There could be a bias in the scheduled OR time when compared with the actual used OR time (such as constantly underestimating the duration of a case to fit it into the schedule) [50]. For instance, consider the case when an OR is scheduled to be open until 3:00PM, but consistently the last patient leaves the room at 3:45PM. If the OR is staffed only until 3:00PM in the schedule, staff will need to work overtime unless there is staff assigned to other ORs available. Hospitals partly manage this variability by staggering shifts of different lengths [108], but the choice of these shifts could be improved if done in systematic way that also considers staff pooling. Another issue with this approach is that the number of required staff depends on the type of procedure, and it can vary even within the same service line and may not be constant throughout the day. Finally, the task of assigning a given shift (and its staff) to a particular service line becomes more complicated if the OR can be shared among multiple service lines.

To overcome these challenges, we propose a two-phase mathematical model for OR long-term staff planning decisions. The first phase is a stochastic model that finds a staffing budget and permanent FTEs per service line that minimize the expected labor cost, using staff pooling and overtime over a planning horizon under different staff hours demand scenarios, to deal with both demand variability and uncertainty. Each day of the planning horizon is divided into shorter *time buckets* (time sub-periods), and the required staff is estimated for each of these time buckets. Then, the second phase of this two-phase model allocates the permanent staff hours to a set of shifts, so that staff is available when required.

The following notation is used in the proposed mathematical formulation.

Sets

Phase I

$S = \{1, \ldots, N_S\}$	Service lines
$W = \{1, \ldots, N_W\}$	Periods (weeks) of the planning horizon
$N = \{1, \ldots, N_N\}$	Demand scenarios

Phase II

$D = \{1, \ldots, 7\}$	Days of the week
$T = \{1, \ldots, b7N_W\}$	Time buckets of the planning horizon; b is the number of time buckets per day
$J = \{1, \ldots, N_J\}$	Shifts
$L = \{8, 9, 10, 12\}$	Full-time shift lengths (hours)
V_t	Set of shift-day pairs (j, d) that cover time bucket t

Parameters

Phase I

X_s^0	Initial number of permanent FTEs for service line s
$F_{s,w,n}$	Demand (staff hours) for period w and service line s in demand scenario n
p_n	Probability of demand scenario n; $\sum_{n \in N} p_n = 1$
H^e	Effective number of hours (spent in the ORs) per budgeted FTE per week
$\alpha_s^p(\alpha_s^o)$	Maximum fraction of pooled (overtime) staff hours of service line s in any period $w \in W$
$\bar{\alpha}_s^p(\bar{\alpha}_s^o)$	Maximum fraction of pooled (overtime) staff hours of service line s averaged over the W periods
$C_s^h(C_s^f)$	Cost of hiring (firing) an FTE for service line s
C_s^r	Cost for an FTE working regular time for service line s
C_s^o	Overtime cost per hour for service line s

Phase II

b	Number of time buckets in a day
B_s	Available (budgeted) permanent FTEs for service line s
H^{std}	Standard number of hours available to schedule (at the hospital) per budgeted FTE per week
H_j^l	Shift j length (hours)
f_l	Weekly frequency of a full-time shift of length l

$R_{s,t}(R_t^{min})$	(Minimum) number of staff required in time bucket t and service line s
$\alpha_s^{lf}(\alpha_s^{pt})$	Maximum fraction of less-than-full-time shifts $H_j^l < 8$ (part-time shifts) hours for service line s
$\Pi_{s,t}$	Penalty for unmet required staff in time bucket t for service line s
Π'_t	Additional penalty for aggregated (i.e., across all service lines) unmet required staff in time bucket t

Variables

Phase I

X_s	Number of permanent FTEs to budget for service line s
$X_s^h(X_s^f)$	Number of permanent FTEs to hire (fire) for service line s
$O_{s,w,n}$	Number of overtime (pooled, slack) staff hours in period w for service
$(P_{s,w,n},$	line s in demand scenario n
$K_{s,w,n})$	

Phase II

$Y_{s,j,d}$	Number of permanent staff (full-time, part-time) assigned to shift j
$(Y^{ft}_{s,j,d},$	on day d, for service line s
$Y^{pt}_{s,j,d})$	
$Z_{s,l}$	Number of full-time permanent staff, with shifts of length l in service line s
$U_{s,t}$	Unmet required staff in time bucket t for service line s
U'_t	Aggregated unmet required staff in bucket time t
$K'_{s,t}$	Staff slack according to the required staff in time bucket t for service line s

The demand ($F_{s,w,n}$) for staff hours is considered on a weekly basis, because with the exception of vacations and absenteeism, each permanent staff member is scheduled every week (about the same hours each week). The number of overtime hours is also computed on a weekly basis (e.g., the number of hours in excess of 40 per week). The effective number of hours per FTE per week (H^e) is the standard number of hours an FTE is scheduled to work per week (e.g., $H^{std} = 40$ hours per week) adjusted to incorporate any time that is not spent in the OR (e.g., lunch time). In Phase I, we consider that for each budgeted FTE (X_s), we have H^e *effective* hours available to set up and staff surgical cases *in* the ORs during a week. In Phase II, we consider that for each available FTE (B_s, where B_s is obtained from X_s if Phase I solution is used in Phase II), we have $H^{std} \geq H^e$ hours available to cover shifts during a week. To analyze the required staff across time, we divide each day of the planning horizon into

b time buckets, considering the required granularity of time (e.g., $b = 48$ half-hour time buckets). Each shift j is defined by a start and end time. Given a full-time shift length l, there is a target weekly frequency (f_l) based on the number of hours per week each full-time staff should be assigned to shifts of length l. For example, if full-time staff should work between 36 and 40 hours per week and could be assigned to 8, 9, 10 or 12-hour shifts, each staff member could do five 8-hour shifts a week, four 9-hour shifts a week, four 10-hour shifts a week, or three 12-hour shifts a week. The required staffing level ($R_{s,t}$) is computed considering the number of open cases at the ORs, for each time bucket t and service line s. However, there could be a required minimum staffing level (R_t^{min}) even when there are no cases at time t, e.g., to respond to emergencies. The staffing structure ($Y_{s,j,d}$) is given for each day d of any week, since the OR block schedule and most scheduling practices are implemented in week-long cycles. The penalty for not meeting the required staffing level ($\Pi_{s,t}$) can vary by time bucket and service line; for example, $\Pi_{s,t}$ may be higher for morning time buckets to ensure that each OR's first case of the day starts on time (since delays would affect the entire day), or lower for time buckets later in the planning horizon. The additional penalty (Π'_t) reflects the policy that staff can be pooled from other services, yet this is undesirable.

2.3.1 Phase I: Staff Budgeting

In Phase I, we have two types of decisions to make. First, we need to decide the number of permanent FTEs to budget for each service line for the planning horizon; and second, we need to decide how we would share these FTEs across the different service lines or use overtime once the staff hours demand is realized. We consider policies that limit the use of staff pooling and overtime for any period of the planning horizon and on average. It is also desired that these policies hold across the different demand scenarios. Therefore, the goal of this first phase is to determine the number

of permanent FTEs to minimize expected labor costs (from regular time and expected overtime), while fulfilling the demand and limiting overtime and staff pooling. This problem is a particular case of two-stage stochastic programming. We assume that the number of demand realizations is finite, or that the demand distribution can be adequately estimated by a finite number of scenarios (although the number of demand scenarios could be very large). Then, it is possible to model this stochastic program as a deterministic optimization problem. The model's Phase I formulation (hereafter also referred as Phase I) is as follows:

$$H^e X_s + O_{s,w,n} + P_{s,w,n} - K_{s,w,n} = F_{s,w,n} \quad s \in S,\ w \in W,\ n \in N \tag{1}$$

$$\sum_{s \in S} P_{s,w,n} \leq \sum_{s \in S} K_{s,w,n} \quad w \in W,\ n \in N \tag{2}$$

$$\frac{1}{N_W} \sum_{w \in W} P_{s,w,n} \leq \bar{\alpha}_s^p [H^e \cdot X_s] \quad s \in S,\ n \in N \tag{3}$$

$$P_{s,w,n} \leq \alpha_s^p [H^e \cdot X_s] \quad s \in S,\ w \in W,\ n \in N \tag{4}$$

$$\frac{1}{N_W} \sum_{w \in W} O_{s,w,n} \leq \bar{\alpha}_s^o [H^e \cdot X_s] \quad s \in S,\ n \in N \tag{5}$$

$$O_{s,w,n} \leq \alpha_s^o [H^e \cdot X_s] \quad s \in S,\ w \in W,\ n \in N \tag{6}$$

$$X_s^0 + X_s^h - X_s^f = X_s \quad s \in S \tag{7}$$

$$X_s, X_s^h, X_s^f, P_{s,w,n}, O_{s,w,n}, K_{s,w,n} \geq 0 \quad s \in S,\ w \in W,\ n \in N \tag{8}$$

$$MIN \quad E(Cost) = \sum_{s \in S} \left[C_s^r X_s + C_s^h X_s^h + C_s^f X_s^f \right] + \sum_{s \in S, w \in W, n \in N} p_n C_s^o O_{s,w,n} \tag{9}$$

Constraints (1) ensure that the demand ($F_{s,w,n}$) is covered with the different sources of labor (permanent staff, overtime, or staff pooling). Constraints (2) limit staff pooling to the available slack labor of the other service lines. Constraints (3)

and (4) limit the average pooling during the planning horizon (across the N_W weeks) and the staff pooling per period (for each week $w \in W$), respectively, as a fraction of the number of permanent staff effective hours for each service line s. Similarly, constraints (5) and (6) limit overtime. Note that constraints (1) to (6) should hold for any demand scenario $w \in W$. Constraints (7) balance the required and the initial numbers of permanent FTEs, to compute the required new hires or lay-offs. All variables are required to be non-negative (8). Notice that the number of permanent FTEs is not necessarily an integer as it is just a standard measure of labor. For example, there could be part-time staff included in the permanent workforce or some full-time staff could work less than the standard FTE hours per week (e.g., 36 rather than 40 hours). Finally, the objective function (9) minimizes the expected total labor cost during the planning horizon, including costs of permanent staff, new hires and lay-offs, and the expected overtime given the set of demand scenarios and the supporting distribution. The expected total staffing cost is equivalent to this expected labor cost, adjusted by the cost of the additional relief FTEs.

2.3.2 Phase II: Staffing Structure

In Phase I, we define X_s, which becomes the number of available FTEs (B_s) for each service line s. In Phase II, we translate these permanent FTEs into shifts by defining a staffing structure to cover the staff requirements by service line s at any time bucket t, while considering the later creation of feasible tours during the medium-term decisions (see Figure 1). The model's Phase II formulation (hereafter also referred as Phase II) is as follows:

$$\sum_{\substack{j \in J, d \in D: \\ (j,d) \in V_t}} Y_{s,j,d} + U_{s,t} - K'_{s,t} = R_{s,t} \quad s \in S,\ t \in T \tag{10}$$

$$\sum_{s \in S} U_{s,t} - \sum_{s \in S} K'_{s,t} \leq U'_t \quad t \in T \tag{11}$$

$$\sum_{\substack{s\in S, j\in J, d\in D:\\ (j,d)\in V_t}} Y_{s,j,d} \geq R_t^{min} \quad t \in T \tag{12}$$

$$\sum_{\substack{j\in J, d\in D:\\ H_j^l \leq 8}} H_j^l Y_{s,j,d} \leq \alpha_s^{lf} \sum_{j\in J,\ d\in D} H_j^l Y_{s,j,d} \quad s \in S \tag{13}$$

$$Y_{s,j,d} = Y_{s,j,d}^{ft} + Y_{s,j,d}^{pt} \quad s \in S,\ j \in J,\ d \in D \tag{14}$$

$$Y_{s,j,d}^{ft} = 0 \quad s \in S, j \in J, d \in D:\ H_j^l < 8 \tag{15}$$

$$\sum_{j\in J, d\in D} H_j^l Y_{s,j,d}^{pt} \leq \alpha_s^{pt}[H^{std} \cdot B_s] \quad s \in S \tag{16}$$

$$Z_{s,l} \geq \frac{1}{f_l} \sum_{\substack{j\in J, d\in D:\\ H_j^l = l}} Y_{s,j,d}^{ft} \quad s \in S,\ l \in L \tag{17}$$

$$Z_{s,l} \geq \sum_{j\in J: H_j^l = l} Y_{s,j,d}^{ft} \quad s \in S,\ l \in L,\ d \in D \tag{18}$$

$$\sum_{j\in J, d\in D} H_j^l Y_{s,j,d}^{pt} + \sum_{l\in L} [l \cdot f_l] Z_{s,l} \leq H^{std} B_s \quad s \in S \tag{19}$$

$$Y_{s,j,d}, Y_{s,j,d}^{ft}, Y_{s,j,d}^{pt}, Z_{s,l} \in \mathbb{Z}^+ \quad s \in S,\ l \in L,\ j \in j,\ d \in D \tag{20}$$

$$U_{s,t}, U_t', K_{s,t}' \geq 0 \quad s \in S,\ t \in T \tag{21}$$

$$MIN \quad Penalty = \sum_{s\in S, t\in T} \Pi_{s,t} U_{s,t} + \sum_{t\in T} \Pi_t' U_t' \tag{22}$$

Constraints (10) compute the gap between the scheduled and required staffing levels for each service line in each time bucket ($U_{s,t}$), which is obtained by balancing the scheduled staff, the unmet staff, and the staff slack on the left-hand side of the constraints (LHS) with the staffing level requirements on the right-hand side of the constraints (RHS). Constraints (11) compute the overall unmet staff on the RHS, considering staff pooling, by adding the unmet staff minus the staff slack of all service lines on the LHS. Constraints (12) ensure that there is a minimum scheduled staffing level in each time bucket. Constraints (13) limit the staff hours from less-than-full-time shifts on the LHS, i.e., shifts less than 8-hours long (which could be less appealing to part-time staff) as a fraction of all the scheduled staff hours for each service line on the RHS. Constraints (14) balance the total number of assigned shifts on the LHS, with the shifts assigned to full- and part-time staff on the RHS, and constraints (15) ensure that only full-time shifts (at least 8-hours long) are assigned to full-time staff. Constraints (16) limit the staff hours scheduled to part-time staff (on the LHS) as a fraction of the available permanent staff hours to schedule (i.e., the standard number of hours to schedule per FTE, H^{std}, times the number of FTEs, B_s, on the RHS).

Constraints (17) make sure that the number of full-time staff working shifts of length l (on the LHS) is greater or equal than the number of full-time shifts of the given length to assign, divided by the number of shifts per week per full-time staff (on the RHS). Constraints (18) make sure that the number of full-time staff on the LHS is greater or equal than the number of shifts to be covered by these staff, at any given day (on the RHS). Then, constraints (17) and (18) give a lower bound on the number of workers $Z_{s,l}$ (assigned to service line s and working shifts of lenght l) needed to cover the scheduled full-time shifts ($Y^{ft}_{s,j,d}$) so that each full-time staff member can be given enough days-off per week according to the shifts' weekly frequency (by working a maximum number of shifts per week), and that no more than one shift per day is assigned per staff member, respectively. Considering the facts that many of the OR

staff works less than 5 days a week and that most of the OR shifts are scheduled during the day [45] and the weekdays [53], these conditions should be in most cases also sufficient (see construction argument by Burns et al. [30]), without compromising minimum rest time between shifts. We assume that all full-time staff $Z_{s,l}$ should be scheduled to only one type of shift length each week (which is preferable in this case). This is similar to the model presented by Bard et al. [20], where staff can only be assigned to shifts that start during a given time window. Appendix A includes a version of constraints (17), (18), and (19), under the alternative assumption that full-time staff can be assigned to shifts of different lengths l in a week.

Constraints (19) add the part-time and full-time scheduled hours on the LHS to ensure they do not exceed the available permanent FTEs hours to schedule by service line on the RHS. Constraints (20) and (21) ensure the integrality and nonnegativity requirements of the variables. Finally, objective function (22) minimizes the penalty resulting from the gaps between the required and the scheduled staffing levels for each and across all service lines.

2.3.2.1 Facilitating Phase II Results Implementation

We consider two characteristics of the staffing structure that is determined in Phase II: (i) How many different shifts compose this staffing structure, and (ii) how different is this staffing structure compared to the current one. Di Gaspero et al. [55] also consider the number of different shifts to cover certain staff requirements and propose a model to minimize it; however, their approach is different as they consider an objective function that penalizes staff excess, shortage, and the number of shifts, but do not consider how these shifts are going to be assigned to staff later on. The number of different shifts is relevant because as this number increases, staff schedules might be harder to manage. Also, it may be easier to implement a staffing structure that is similar to the current one, as it would require fewer changes with respect to

the current schedules. To give more control on the proposed staffing structure, we introduce the following constraints, solved in addition to those described in Section 2.3.2:

Phase II: Additional Parameters

M	Maximum number of staff that can be assigned to any shift j on any day d for any service line s
N_J^{max}	Maximum number of different shifts that can be assigned
$Y_{s,j,d}^0$	Current number of staff assigned to shift j on day d for service line s
Π^q	Penalty from deviating from initial staffing structure $Y_{s,j,d}^0$

Phase II: Additional Variables

γ_j	1 if shift j is included in the staffing structure $Y_{s,j,d}$; 0 otherwise
$Q_{s,j,d}^+(Q_{s,j,d}^-)$	Positive (negative) deviation between the current staffing structure $Y_{s,j,d}^0$ and the proposed staffing structure $Y_{s,j,d}$

$$Y_{s,j,d} \leq M \cdot \gamma_j \quad s \in S,\ j \in J,\ d \in D \tag{23}$$

$$\sum_{j \in J} \gamma_j \leq N_J^{max} \tag{24}$$

$$Y_{s,j,d}^0 + Q_{s,j,d}^+ - Q_{s,j,d}^- = Y_{s,j,d} \quad s \in S,\ j \in J,\ d \in D \tag{25}$$

$$\gamma_j \in \{0,1\} \quad j \in J \tag{26}$$

$$Q_{s,j,d}^+, Q_{s,j,d}^- \geq 0 \quad s \in S,\ j \in J,\ d \in D \tag{27}$$

$$MIN \quad Penalty' = Penalty + \Pi^q \sum_{\substack{s \in S, j \in J, \\ d \in D}} \left[Q_{s,j,d}^+ + Q_{s,j,d}^-\right] \tag{28}$$

Constraints (23) ensure that if shift j is included in the staffing structure ($Y_{s,j,d}$), then $\gamma_j = 1$. Constraint (24) limits the number of shifts that can be included in the staffing structure to N_J^{max}. Constraints (25) compute the difference between the resulting staffing structure and the current one, by balancing the current staffing structure ($Y^0_{s,j,d}$) plus/minus the positive/negative deviation on the LHS and the recommended staffing structure ($Y_{s,j,d}$) on the RHS. Constraints (26) and (27) ensure the binary and non-negativity requirements of the variables, respectively. Finally, the adjusted objective function (28) adds the additional penalty that results from the deviations in (25) to the original objective function (22).

2.4 Case Study

We implement Phases I and II with input generated using the hospital's surgical data, including: type of patient (outpatient, inpatient), surgeon, anesthetist, procedure description, surgical service, and the *case time stamps* (case's scheduled start and end times, patient's OR wheels-in and wheels-out times, and anesthesia start and end times). Next, we describe how we compute the demand for staff hours, the staffing levels, and other parameters; and we describe Phases I and II results and compare them with those obtained with the hospital's current practices. We discuss additional implementation parameters for Phases I and II in Appendix B.

2.4.1 Demand Data Analysis

The first step in the demand data analysis is to classify each surgical case into one of the three service lines: cardio/vascular, neuro/ortho and general. The second step is to estimate the demand for staff hours for each service line. We assume that once an OR is open for its first case, it generally remains staffed until its last case is completed. This is consistent with the current practice, as the OR staff is often busy between cases preparing the room for the next case (*room turnover*). If a case is followed by a case of a different surgical service, we assume that the time between cases is staffed

according to the staffing requirements of the following case. We assume a limit of 90 minutes for room turnover between cases, starting at the wheels-out time of the previous case. This is long enough to cover for an OR cleaning and set up, even for those ORs with reputation of slow turnover times [47]. In our data, the time between two consecutive cases in an OR is less than 90 minutes in more than 80% of the cases. As the current practice, we assume that there are 30 minutes to prepare an OR for its first case before the scheduled start time, and 30 minutes to clean it after the wheels-out time of its last case (in a 2 months time study, cleaning time took 30 minutes or less in 95% of the cases). We use the case wheels-out time rather than the scheduled end time since given some scheduling practices and surgeons' behavior, case duration is commonly underestimated to fit a case in a given OR's schedule [50]. If staff are not scheduled for the entire duration of the day's surgeries in an OR, overtime will be required and/or the case could be delayed. See Figure 2 for an example of staff hours demand calculation. There is an orthopaedic case followed by a cardio case, with one hour between the two cases, and the cardio case lasts (wheels-in to wheels-out) two hours, i.e., the case requires three hours of OR time including room turnover. Cardio and orthopaedic cases require two scrub techs and one circulator (other cases require one of each). Therefore, the cardio/vascular line requires three circulator hours and six scrub tech hours for this cardio case.

In this hospital, there is one night shift team composed of one circulator and two scrub techs that is available in case of an emergency outside the ORs regular hours. We identify those cases that start after the ORs regular hours and exclude them from the staff hours demand calculation. If there are more than one of these cases at the same time, we only exclude the case that starts first. The reason for doing this is to compute the required staff hours demand for the cases during the ORs regular hours, and add the corresponding night team staff hours without double counting them.

The third step in the demand data analysis is to compute the staffing levels by

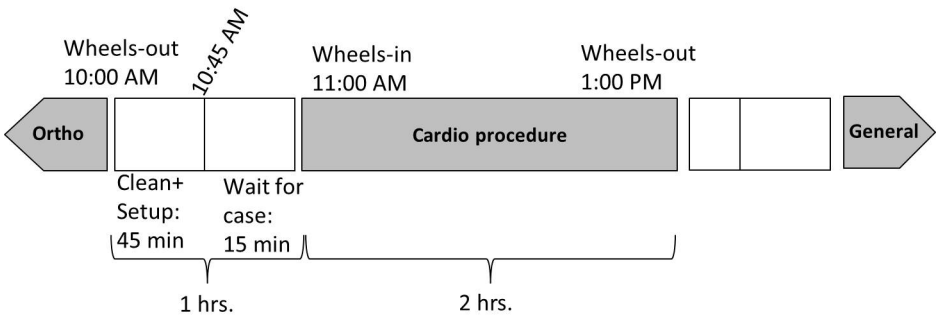


Figure 2: An example of a case staff hours demand calculation.

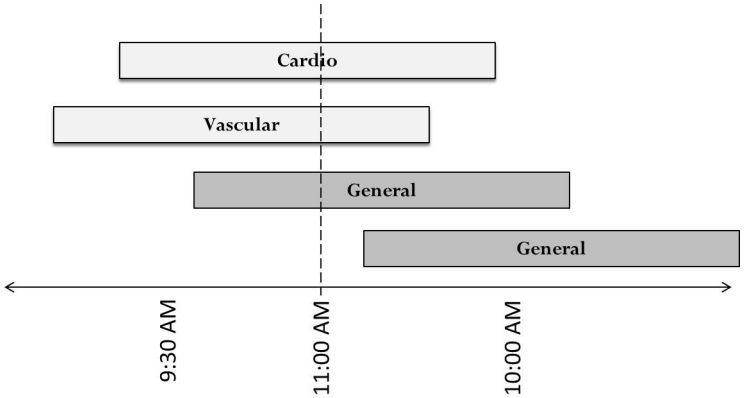


Figure 3: An example of a case staff demand calculation by time bucket.

day and time bucket. We divide each day into $b = 48$ time buckets of 30 minutes each. We round each case time stamp to the closest time bucket. We calculate the number of staff required during each bucket on each day, for each service line. We consider that staff are required when the patient is in the OR and during the room turnover. For example, in Figure 3 at the time bucket corresponding to 11:00 AM, there is a need for three scrub techs for the cardio/vascular service line (two scrub techs for the cardio case and one for the vascular case), and one scrub tech for the general service line.

2.4.2 Forecasting Staff Hours Demand

To make Phase I budgeting decisions, we need to forecast staff hours demand during the planning horizon. In this case, the planning horizon is the first 48 weeks of 2012,

with a forecast made about six months in advance. A demand forecast for a specific week within such a distant time horizon is likely to be inaccurate. Nevertheless, we are not interested in estimating the exact demand for each week of the coming year. We need to make a decision on the number of permanent FTEs to budget for each service line for the planning horizon, and then manage demand variability on a weekly basis through the use of overtime or staff pooling among services. Therefore, we are only interested in the weekly *demand distribution* during the planning horizon. We propose the following procedure:

1. Using a linear regression, estimate the expected staff hours demand, $E(F_{s,w})$, for each service line s and each week w in the planning horizon.

2. Make adjustments to the fitted models to incorporate additional information not reflected in the historical data (for instance, the retirement, long-term absence, or arrival of surgeons, and the closing or opening of nearby surgical facilities).

3. Use the standard error (SE) of the fitted models to add a random deviation to the adjusted expected demand, for each service line s and each w week in the planning horizon, i.e.,

$$F_{s,w,n} = E(F_{s,w}) + \epsilon_{w,s,n}, \tag{29}$$

   where $\epsilon_{w,s,n} \sim N(0, SE_s^2)$, and $F_{s,w,n}$ is the staff hours demand for period w and service line s in demand scenario n.

4. Repeat step 3 to generate $N' = \{1, \ldots, N_{N'}\}$ demand scenarios for each week of the planning horizon.

Dexter et al. [52] also use the SE of the forecasting error to generate confidence intervals for the OR staff hours demand for a given surgical service. They predict the demand for the next four-week period using 12 previous four-week periods using

moving average time series. We are interested in longer time horizons (e.g, a year long weekly prediction, done months in advance), so we consider factors that could affect the demand to isolate an increasing or decreasing demand trend. We consider the following three demand predictors for each of the six regression models (for each service line and staff type), based on the OR manager suggestions and our analysis of historical data: (i) time (i.e., week period), (ii) if there are holidays during the week, and (iii) if the week falls in a month with higher than average workload. Moore et al. [95] also relate some of the surgical demand variance with holidays in their time series analysis of daily surgical demand. They also find week day cycles and 1-day and 7-days lag effects. Initially, we also considered lag effects in the weekly surgical demand, but we did not find that these factors were statistically significant in our case. On the other hand, Dexter et al. [54] do not find evidence of systematic month-to-month variation in their analysis of the National Survey of Ambulatory Surgery. This means that factor (3) does not generalize to other institutions. However, they note that some seasonal variation (specifically, in the afternoon) could be the case for some surgical groups and suggest checking for seasonality. We discuss the details of the regression model and the assumptions about the normality of the error in Section 2.4.2.1.

While Dexter et al. [90] also recommend the use of historical data and simple statistical methods to predict future surgical demand over local demographic data, they also recognize limitations to incorporate other less quantitative factors such as changes in competition. We suggest in the second step of the procedure above, to consider relevant additional information to adjust the demand predictions when possible.

A comparison of the staff hours weekly demand distributions resulting from the proposed procedure and the actual demand distributions during the first 48 weeks of 2012 does not show statistically significant differences. However, there are statistically

significant differences when historical demand data are used directly as estimator of future demand distributions. We present the details of these results in Section 2.4.2.2.

2.4.2.1 A Regression Model for Forecasting Demand for ORs Staff Hours

We propose the following linear regression model to forecast the expected staff hours demand, $E(F_{s,w})$, for service lines s and week w:

$$E(F_{s,w}) = K_s + HD_s \cdot w_{HD} + HM_s \cdot w_{HM} + TT_s \cdot w_{TT} \tag{30}$$

HD_s is the regression coefficient to factor for holidays, and w_{HD} equals to 1 if there is a hospital official holiday during week w. HM_s is the coefficient to factor for a month of a demand higher than average for service s and week w, and w_{HM} is the percentage of days of week w that fall in a month with higher demand. TT_s is a coefficient to account for an increasing or decreasing staff demand trend with respect of time, which can result from changes in the volume and/or mix of surgical cases, and w_{TT} is the week number that results from enumerating the first week of the data set as 1, the second as 2, and so on. K_s is the regression constant.

In this hospital, staff planning decisions are made as early as July for the planning horizon starting in January of the next year and ending in December. We use two years of data, from week 28 in 2009 to week 27 in 2011, to estimate the demand for staff hours in 2012 (weeks 1 to 48 in 2012, as data from the last weeks of 2012 are not available for comparison). We do not use data from week 28 and after in 2011, since we need to estimate 2012 demand in July of 2011. In addition, we eliminate the first week of 2010 and 2011 from the data because the volume is too low as these weeks include only three and two weekend days respectively, i.e., most of the cases of that first week belong to 2009 and 2010, respectively. We use these 104 weeks of available historical data and Minitab 16 for the regression analysis.

We run a stepwise regression and all the proposed predictors are retained (p-value

Table 1: Regression models results.

Staff Type	Service line	K		HD		HM		TT	
		Coef.	p-value	Coef.	p-value	Coef.	p-value	Coef.	p-value
Circulator	Cardio/vascular	132.409	<0.001	-36.556	<0.001	11.79	0.078	-0.08812	0.343
	Neuro/ortho	174.276	<0.001	-49.594	<0.001	12.408	0.016	0.17622	0.021
	General	195.313	<0.001	-43.851	0.006	8.167	0.114	-0.115	0.105
Scrub tech	Cardio/vascular	233.45	<0.001	-54.21	0.003	22.85	0.057	-0.3306	0.043
	Neuro/ortho	279.177	<0.001	-82.49	<0.001	25.37	0.014	0.3279	0.008
	General	195.313	<0.001	-43.851	0.006	8.167	0.114	-0.115	0.105

Table 2: Regression models statistics.

Staff Type	Service line	SE	R^2	Regression p-value	Error AD p-value	Error Lag-1 LQB p-value
Circulator	Cardio/vascular	26.938	15.2%	0.001	0.525	0.203
	Neuro/ortho	22.9703	31.9%	<0.001	0.764	0.751
	General	21.596	25.1%	<0.001	0.941	0.920
Scrub tech	Cardio/vascular	48.61	16.8%	<0.001	0.110	0.413
	Neuro/ortho	37.1	32.0%	<0.001	0.913	0.559
	General	21.596	25.1%	<0.001	0.941	0.920

< 0.15) for both circulators and scrub techs and all service lines. The only exception is the case of circulators for the cardio/vascular service, where TT_s has low significance (p-value > 0.15); however, given the importance of capturing time trends, the fact that the coefficient is very small, and to have consistency throughout the models, we keep the week number as a predictor in all the six regression models. Table 1 shows the results for the coefficients and respective p-values. The results for circulators and scrub techs are the same for the general service line, since the scrub tech and circulator requirements are the same for this service line.

Table 2 shows the standard error (SE), the coefficient of determination (R^2), and the regression's p-values. The R^2 values are low for these models (15-32%). However, we only use the regression results as an estimation of the expectation of the demand in each week, and then incorporate demand variability using an error term. To test the distribution of the regression error, $\epsilon_{w,s,n}$, we generate the expected staff hours demand for each of the periods of the baseline data, using the results in Table 1. Then, we compare the regression estimation and the actual demand value for each week; the difference is the error term of the regression. The mean of the

error is almost zero ($< |0.001|$ for all cases) and the standard deviation is the SE (see Table 2). We test the null hypothesis H_0 that the error follows a normal distribution using normal probability plots and the Anderson-Darling (AD) test. From the AD p-values (> 0.50 for all cases but one, where it is 0.11, see Table 2), there is not enough evidence to reject H_0. We also test the independence of the errors using autocorrelation plots and the Ljung−Box Q (LBQ) test for lags up to: 1 week, 4 weeks (a month), and 26 weeks (around 6 months). From the corresponding LBQ p-values (> 0.20 for all cases), there is not enough evidence to reject the null hypothesis that the autocorrelations are different from zero up to the given lag. See Table 2 for the corresponding LBQ p-values for one-week lag. In addition, we do not find evident violations of the homoscedasticity assumption in the regression errors plots vs. time and vs. predicted value. Hence, we assume that the regression error follows a normal distribution with mean zero and a standard deviation equal to SE. Under these assumptions, we can use the distribution of the regression error to add a random error term to the expectation to obtain a realization of the future weekly demand that incorporates both the expectation of the demand and its variability as expressed in equation (29).

2.4.2.2 Evaluating ORs Staff Hours Demand Distribution Forecast

Following the procedure described in Section 2.4.2, we generate scenarios for the staff hours demand forecast for each service s and week w of the planning horizon ($F_{s,w,n}$). We are interested in analyzing the accuracy of this estimation by comparing it to the actual demand distribution in 2012.

Since our aim is to compare two demand distributions and not specific demand realizations for a particular week, we do not use the traditional measures to evaluate forecast accuracy such as the mean absolute percentage error (MAPE). In Figures 4 and 5, we show the empirical distributions of: most recent 52 weeks of historical

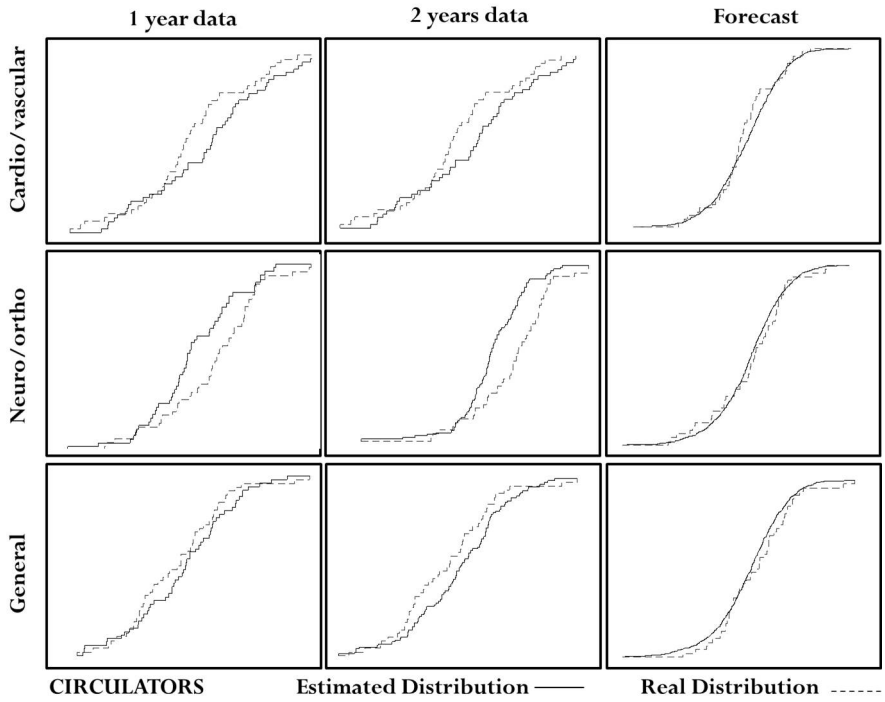


Figure 4: Estimated and actual weekly demand empirical distribution functions for circulators.

data (from week 28 in 2010 to week 27 in 2011, with the exception of the first week of 2011), 104 weeks of historical data (from week 28 in 2009 to week 27 in 2011, with the exception of the first week of 2010 and 2011), and the demand forecast distribution derived from the proposed procedure (and the available 104 weeks of historical data); and we compare them to the actual demand empirical distribution during the planning horizon (weeks 1 to 48 in 2012). We visually observe that the demand forecast distribution curve is closer to the actual distribution than the demand distribution of historical data for both circulators and scrub techs and all service lines. With the regression model, we are able to incorporate demand time trends and characteristics of the planning horizon (holidays, months of higher demand).

To quantify the estimation accuracy for our proposed demand distribution, we compute the Kolmogorov-Smirnov (KS) statistic [120]. The two-sample KS test is one of the commonly used non-parametric tests to compare two empirical distributions

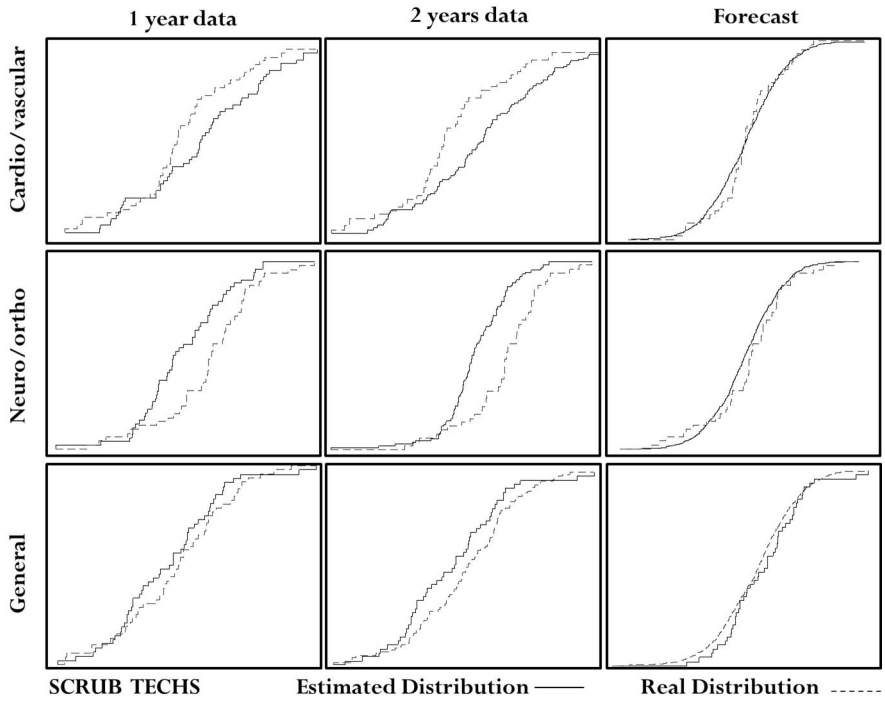


Figure 5: Estimated and actual weekly demand empirical distribution functions for scrub techs.

obtained from two different data sets, because it is sensible to both location and shape differences. The KS statistic measures the maximum distance between the two distributions, which is compared with a critical value for a given confidence level. The KS statistics are smaller when using the proposed procedure to estimate the demand distribution during the planning horizon (see Table 3). When the KS statistics are compared to the critical value, there is not enough evidence (at a 95% confidence level) to conclude that the underlying distributions are different when using the proposed demand forecasting procedure.

Finally, with the proposed procedure we can add robustness to Phase I results by generating and considering several demand scenarios, rather than just a number of scenarios that can be derived directly from the available historical data.

Table 3: Kolmogorov-Smirnov (KS) statistic and critical value (CV) at a α=0.05, for each method.

Staff Type	Service line	52 weeks data (CV=0.272)	104 weeks data (CV= 0.237)	Forecast (CV=0.199)
Circulator	Cardio/vascular	0.248	0.293	0.144
	Neuro/ortho	0.302	0.344	0.101
	General	0.133	0.171	0.111
Scrub tech	Cardio/vascular	0.286	0.354	0.175
	Neuro/ortho	0.360	0.447	0.160
	General	0.133	0.171	0.111

2.4.3 Phase I Results

Following the procedure described in Section 2.4.2, we can generate $N' = \{1, \ldots, N_{N'}\}$ independent and identically distributed (i.i.d.) random demand scenarios for the planning horizon. Under the assumption that this random sample of demand scenarios successfully estimate the true distribution of the demand during the planning horizon (see Section 2.4.2.2), it can be shown that as $N_{N'} \to \infty$, the approximation of Phase I by using these N' i.i.d demand scenarios with probability $p_n = 1/N_{N'}$, $n \in N'$, converges exponentially fast to the original problem with all N demand realizations with probability p_n, $n \in N$. Moreover, under the assumption of a discrete distribution of demand, Monte Carlo sampling based methods (as the one proposed) can be very efficient, and a relatively small $N_{N'}$ can give good results (see [107] for more details on the convergence of Monte Carlo approximations of stochastic programs).

Using CPLEX, we solve Phase I with $N_{N'} = 30$ forecasted demand scenarios for each of the 48 weeks of the planning horizon. It solved almost instantaneously (less than one second). The hospital did not have established staff pooling and overtime policies; therefore, for comparison we determine a staffing policy for which the total number of permanent FTEs employed by the hospital during the planning horizon is approximately the same as Phase I results (under the assumptions discussed in Appendix B), for both circulators and scrub techs. However, the allocation of these FTEs across the three service lines differs between the results of Phase I and the hospital's CP (see Table 4). While the FTEs dedicated to the neuro/ortho service line

Table 4: Difference between Phase I results (for the number of FTEs by service line and staff type) and the hospital's current practice (CP).

Service Lines	Circulators	Scrub Tech	Total
Cardio/Vascular	16.3%	7.4%	11.1%
Neuro/Ortho	-2.6%	1.0%	-0.5%
General	-15.0%	-9.2%	-12.4%
Total	-2.9%	-0.0%	-1.3%

remained approximately the same, Phase I results recommend increasing the FTEs dedicated to the cardio/vascular service line and decreasing the FTEs for the general service line, for both circulators and scrub techs. The ratio between the hospital's CP cardio/vascular FTEs and general FTEs is close to the ratio of their average demands, but the demand coefficient of variation (c.v.) of the cardio/vascular service line is larger than the demand c.v. of the general service line (for both circulators and scrub techs). Phase I results suggest a compensation for this larger variation by increasing the allocated FTEs.

We compute the resulting overtime and staff pooling hours during the 48-week planning horizon to evaluate the permanent FTEs budget by service line using the observed demand during that period of time. We assume that if the demand for the week is greater than the FTEs dedicated to a particular service line, staff dedicated to other service lines could be used if available, and overtime is used otherwise (at a higher cost). This is consistent with the hospital's CP. We compare Phase I budget results with the hospital's. Using the Phase I budget, scrub techs pooling is slightly reduced with the redistribution of FTEs without increasing the budget (see Table 5). Since Phase I and the hospital's permanent FTEs budget is about the same, there is no difference in staff overtime.

We also evaluate both Phase I and the hospital's CP budgets under other potential demand scenarios for the planning horizon. Using Minitab 16, we fit a Logistic distribution (which resembles a Normal distribution but with heavier tails) to the available 48 weeks of 2012 historical staff demand data for the scrub techs of the

Table 5: Number (percentage) of staff pooling hours given the hospital's current practice (CP) and the proposed Phase I budgets, under the planning horizon's actual demand and estimated alternative demand scenarios.

Staff Type	Budget	Actual Demand	Alternative Demand Scenarios
Circulator	CP	0 (0.00%)	6 (0.02%)
	Phase I	0 (0.00%)	3 (0.01%)
Scrub tech	CP	186.4 (0.45%)	229.9 (0.54%)
	Phase I	165.8 (0.40%)	192.1 (0.45%)

neuro/ortho service line, and a Normal distribution for the other staff types and service lines. For the Logistic distribution the AD p-value=0.18 and for the Normal distributions AD p-values ≥ 0.14, i.e., there is not enough evidence to reject the null hypotheses that our demand data of 2012 follow these distributions. We generate 30 demand scenarios for each of the 48 weeks of the planning horizon with the fitted 2012 demand distributions for each OR staff type and service line. Average staff pooling under the alternative demand scenarios is moderately reduced, particularly for scrub techs (see Table 5). This means that the proposed budget is robust under other potential demand scenarios for 2012.

2.4.3.1 Effects of the Staff Pooling and Overtime Policies on the Staff Budget

We solve Phase I under different parameter settings for the maximum average overtime ($\bar{\alpha}_s^o$) and pooling ($\bar{\alpha}_s^p$). Figures 6 and 7 show that low levels of overtime and staff pooling are sufficient to reduce the total number of permanent FTEs required for both circulators and scrub techs. Staff pooling is helpful only if there are staff available in other service lines. Hence, at low values of $\bar{\alpha}_s^p$, the number of FTEs decreases fast for circulators and particularly for scrub techs, and then remain steady. Also, as more overtime is allowed, the minimum number of FTEs is reached at higher levels of staff pooling as the number of FTEs are reduced and staff pooling becomes more important. Similar to staff pooling, there are diminishing returns with overtime, and in particular, when $\bar{\alpha}_s^o > 0.1$ allowing more overtime does not make a difference.

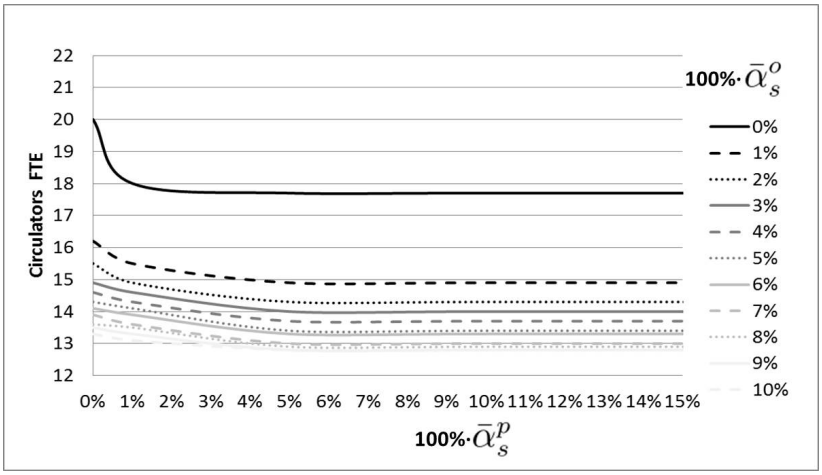


Figure 6: Policies on maximum average pooling ($\bar{\alpha}_s^p$) and overtime ($\bar{\alpha}_s^o$) effect on the FTEs budget for circulators.

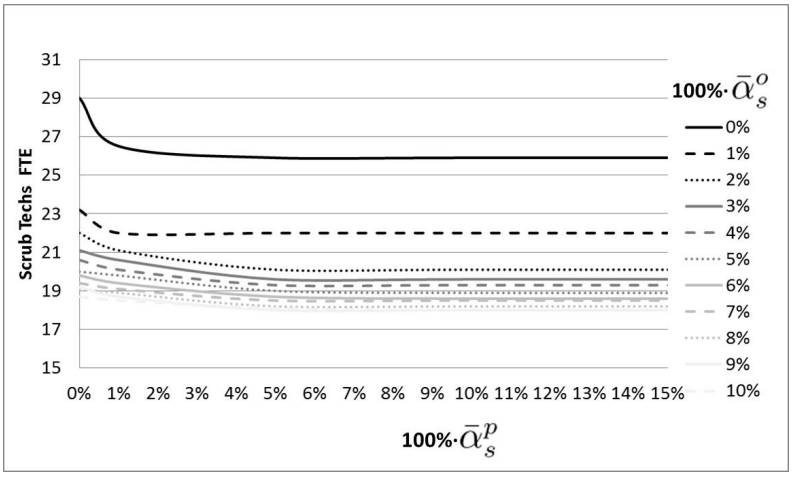


Figure 7: Policies on maximum average pooling ($\bar{\alpha}_s^p$) and overtime ($\bar{\alpha}_s^o$) effect on the FTEs budget for scrub techs.

Table 6: Phase II settings for the staff budget B_s and the maximum number of different shifts N_J^{max}.

	N_J^{max} under CP	Flexible N_J^{max}
B_s under CP	(1)	(2)
B_s from Phase I	(3)	(4)

2.4.4 Phase II Results

We solve Phase II as described in Sections 2.3.2 and 2.3.2.1 (constraints (10)--(21) and (23)--(27), and objective function (28)), under four different settings: (1) the number of permanent FTEs (B_s) and the number of shifts correspond to the hospital's CP according to the staffing structure in December 2012, (2) hospital's permanent FTEs under CP but with no restriction on the maximum number of different shifts (i.e., $N_J^{max} = N_J$), (3) Phase I budget and the hospital's CP on the maximum number of different shifts, and (4) Phase I budget and no restriction on the number of different shifts. In all these settings, we assume that the penalty for deviating from the current staffing structure is small in comparison to the penalties for unmet required staff (where, $\Pi^q = 0.001\Pi = 0.002\Pi'$). These four settings are shown in Table 6. We use one year of historical demand data (week 28 in 2010 to week 27 in 2011) as an estimate for the staffing levels during the planning horizon (i.e, the first 48 weeks of 2012).

Figure 8 shows the percentiles of the number of scrub techs required during each time of the week for the planning horizon. These staffing levels are contrasted with those that result from the hospital's CP staffing structure and Phase II results under setting (1). Compared to CP, the Phase II staffing structure better matches the required staff patterns by redistributing and staggering work shifts. For example, in the morning and early afternoon on Mondays, Phase II results better follow the staff demand patterns than the hospital's CP staffing structure, which exceeds the demand. Similarly, in the late morning and early afternoon on Wednesdays, Phase II results better cover potential demand peaks.

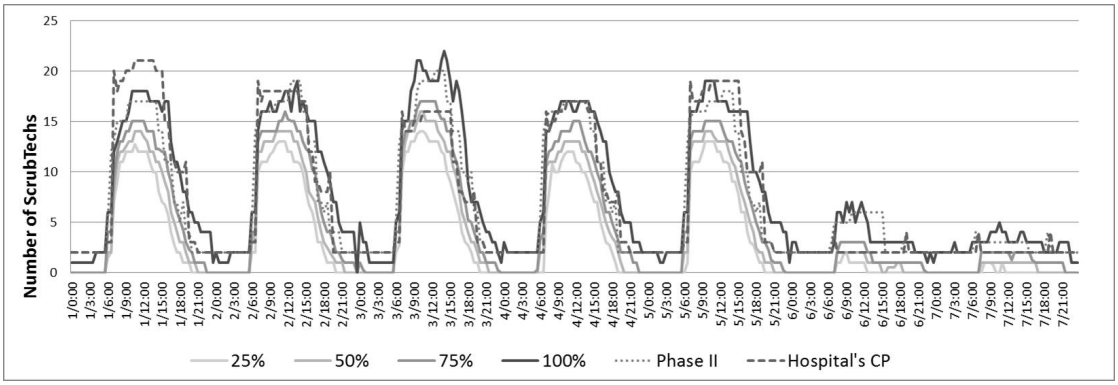


Figure 8: Comparison of the scrub techs staffing levels resulting from the Phase II recommended staffing structure with setting (1) and the hospital's current practice (CP) staffing structure vs. the observed percentiles of the required number of scrub techs during the planning horizon (weekday number/time of the day, where Monday=1).

In Table 7 we report the Phase II optimal *Penalty* value for the objective function (22) for each of the four settings in Table 6, for circulators and scrub techs, as well as the CPLEX running time. Allowing any number of different shifts (or a sufficiently large N_J^{max}) reduces *Penalty* by about 4-5% for circulators and 3-4% for scrub techs. Also, based on Phase II staffing levels input data, there is an advantage of using the FTEs budget as recommended in Phase I (about 6-7% *Penalty* reduction for circulators and 4-5% for the scrub techs). However, *Penalty* is computed based on the input data of the historical staffing levels observed from week 28 in 2010 to week 27 in 2011, which correspond to part of the surgical data also used as input to forecast demand scenarios in Phase I. We are interested in evaluating Phase II results with respect to the actual staffing levels when the staffing structure would be implemented. We define $t \in T'$ as the set of time buckets that correspond to the planning horizon, i.e., the first 48 weeks of 2012. For a given service line s, we measure the gap between the required staffing levels during the planning horizon and those resulting from the staffing structure as follows:

Table 7: Phase II settings, optimal *Penalty*, and running time (minutes).

Staff Type	Budget (B_s)	Max. Different Shifts (N_J^{max})	*Penalty*	Time
Circulators	CP	CP	1488	88.8
		Flexible	1424	0.7
	Phase I	CP	1394	127.5
		Flexible	1335	1.2
Scrub Techs	CP	CP	5288	519.0
		Flexible	5081	0.7
	Phase I	CP	5021	310.1
		Flexible	4863	3.9

$$Gap_s\% = \frac{\sum_{t\in T'} \Pi_{s,t} \max\{R_{s,t} - \sum_{\substack{j\in J, d\in D: \\ (j,d)\in V_t}} Y_{s,j,d}, 0\}}{\sum_{t\in T'} \Pi_{s,t} R_{s,t}} \cdot 100\% \tag{31}$$

$Gap_s\%$ does not consider the option of staff pooling. We compute a gap considering staff pooling, adding up the required staff for all service lines and comparing this with all the available staff.

$$Gap\% = \frac{\sum_{t\in T'} \Pi'_t \max\{\sum_{s\in S} R_{s,t} - \sum_{\substack{s\in S, j\in J, \\ d\in D:(j,d)\in V_t}} Y_{s,j,d}, 0\}}{\sum_{t\in T', s\in S} \Pi'_t R_{s,t}} \cdot 100\% \tag{32}$$

Table 8 shows the results for $Gap\%$, $Gap_s\%$, and $\sum Gap_s\%$ (as an overall gap sum that does not considers staff pooling), given the hospital's and Phase II staffing structures (resulting from each of the four settings given in Table 6), evaluated based on the actual staffing levels during the planning horizon. The Phase II staffing structure achieves better results than the hospital's for all settings. However, a larger maximum number of different shifts N_J^{max} does not seem to consistently improve the gaps as discussed in Section 2.4.4.1.

2.4.4.1 Adding Additional Constraints to Facilitate Phase II Results Implementation

To analyze the impact of constraint (24), we solve Phase II using the hospital's FTEs budget and gradually increasing the maximum number of shifts N_J^{max}. Figures 9

Table 8: Hospital's current practice (CP) and Phase II staffing structures $Gap\%$ and $Gap_s\%$ for the first 48 weeks of 2012 for different Phase II settings. $s = 1, 2, 3$ corresponds to cardio/vascular, neuro/ortho and general service line, respectively.

Staff Type	Budget (B_s)	Max. Different Shifts (N_J^{max})	$Gap\%$	$Gap_1\%$	$Gap_2\%$	$Gap_3\%$	$\sum Gap_s\%$
Circulators	*CP staffing structure*		*1.39%*	*14.57%*	*4.81%*	*7.67%*	*27.05%*
	CP	CP	0.45%	10.19%	3.49%	1.02%	14.70%
		Flexible	0.32%	9.84%	3.55%	1.08%	14.47%
	Phase I	CP	0.19%	6.61%	3.87%	2.10%	12.58%
		Flexible	0.43%	7.93%	3.50%	2.10%	13.53%
Scrub Techs	*CP staffing structure*		*2.54%*	*18.49%*	*13.01%*	*3.59%*	*35.09%*
	CP	CP	1.33%	13.45%	8.32%	2.12%	23.89%
		Flexible	1.22%	13.48%	8.43%	1.84%	23.75%
	Phase I	CP	1.32%	12.22%	8.50%	2.20%	22.92%
		Flexible	1.22%	12.33%	8.18%	2.63%	23.14%

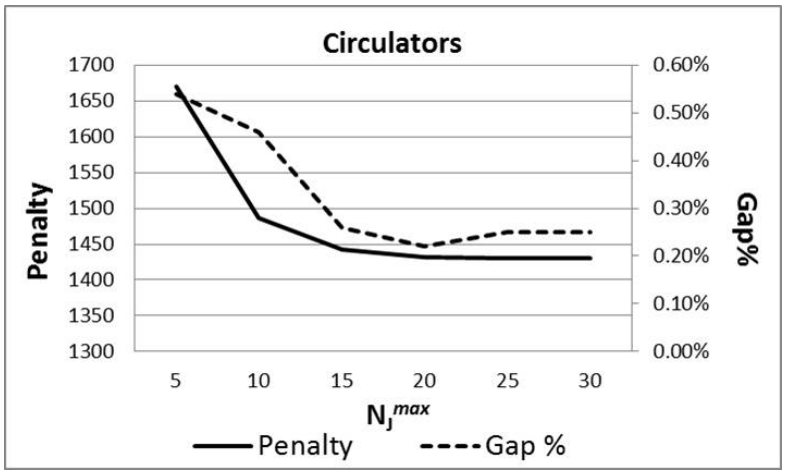


Figure 9: Circulators *Penalty* and $Gap\%$ vs. the maximum number of different shifts N_J^{max}.

and 10 show the Phase II results for *Penalty* and $Gap\%$ for circulators and scrub techs, respectively. *Penalty* decreases as the number of shifts increases, while it is not always the case for $Gap\%$, because $Gap\%$ is computed based on the actual staffing levels observed during the planning horizon, rather than the estimations used as Phase II input.

To analyze the effect of constraints (25), we solve Phase II using the hospital's FTEs budget and N_J^{max}, gradually increasing Π^q, assuming the OR manager wants to get a staffing structure similar to the current one. As shown in Figures 11 and 12, *Penalty* and $Gap\%$ increase as Π^q increases, until Π^q is large enough ($\Pi^q > 500$)

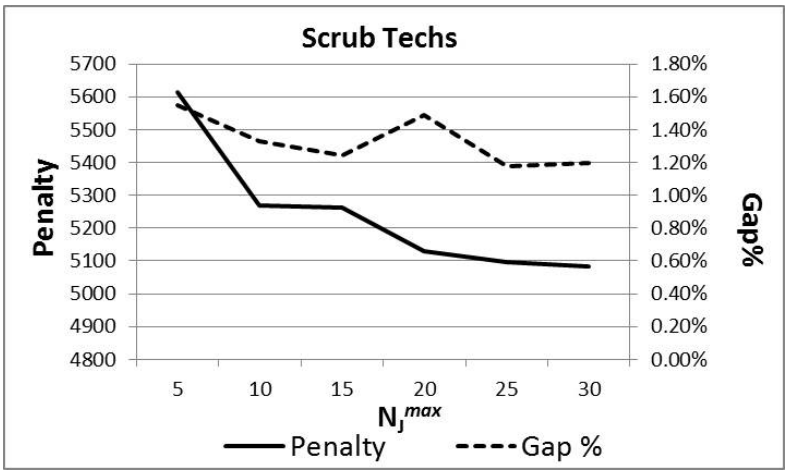


Figure 10: Scrub Techs *Penalty* and *Gap*% vs. the maximum number of different shifts N_J^{max}.

where Phase II results converge to the current staffing structure $Y^0_{s,j,d}$.

2.4.4.2 Phase II: Heuristic

Phase II settings that limit N_J^{max} in Table 7 can take more than couple hours to solve. In Figure 13, we observe that as N_J^{max} increases, the running time to solve Phase II decreases. We also observe in Figure 14 that as we increase Π^q, while keeping the hospital's CP N_J^{max}, the running time tends to decrease as $Y_{s,j,d}$ converge to $Y^0_{s,j,d}$. These observations motivate the heuristic we describe next.

The first step of the heuristic is to solve Phase II relaxing the integrality requirements of the number of shifts and staff: $Y_{s,j,d}, Y^{ft}_{s,j,d}, Y^{pt}_{s,j,d}$, and $Z_{s,l}$. γ_j $j \in J$, remain as binary variables since we want to select at most N_J^{max} different shifts. We use the potentially fractional solution for $Y_{s,j,d}$ as the 'current staffing structure' $Y^0_{s,j,d}$, and we solve the original Phase II formulation with all the integrality constraints with a given Π^q. We use the four settings for B_s and N_J^{max} in Table 6, with three different values for Π^q: 100, 500, and 1000 (for reference, $\Pi_{s,t}$=1 for all $s \in S$ and $t \in T$).

In Table 9, we show the ratio of the *Penalty* resulting from the heuristic and the optimal solutions, and the total running time for the heuristic. The results with $\Pi^q = 1000$ and $\Pi^q = 500$ are similar, with a *Penalty* of the heuristic solution 2-32%

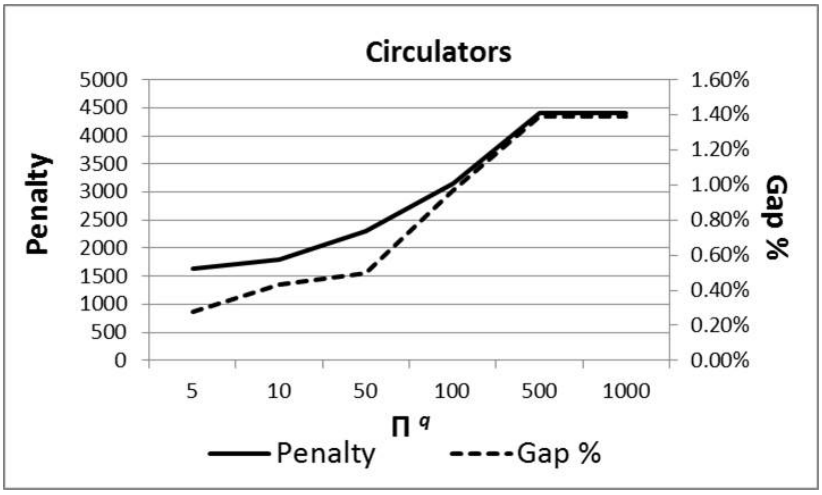


Figure 11: Circulators *Penalty* and *Gap*% vs. penalty Π^q for deviating from current staffing structure $Y^0_{s,j,d}$.

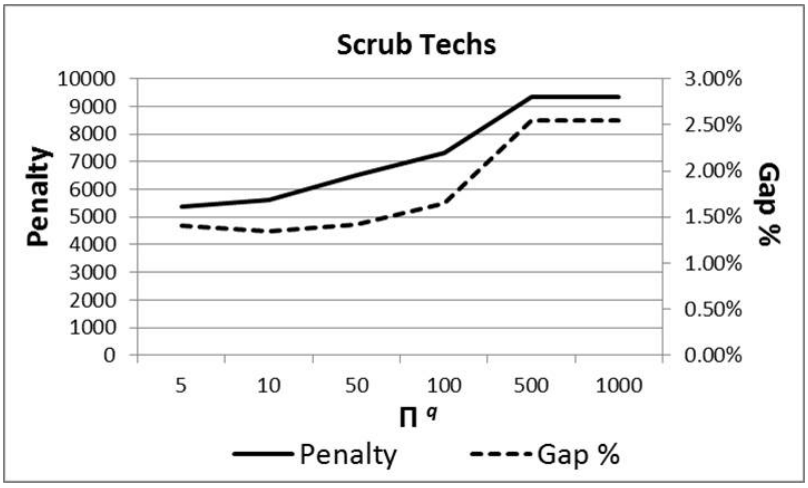


Figure 12: Scrub techs *Penalty* and *Gap*% vs. penalty Π^q for deviating from current staffing structure $Y^0_{s,j,d}$.

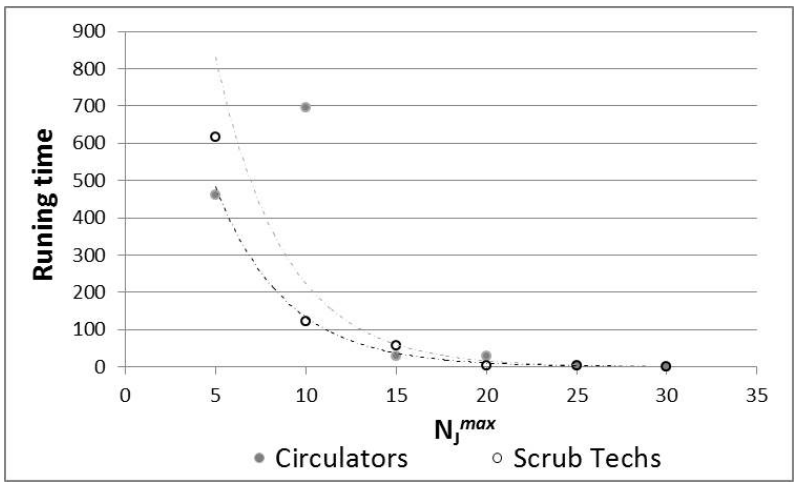


Figure 13: Circulators and scrub techs running time (minutes) vs. the maximum number of different shifts N_J^{max}.

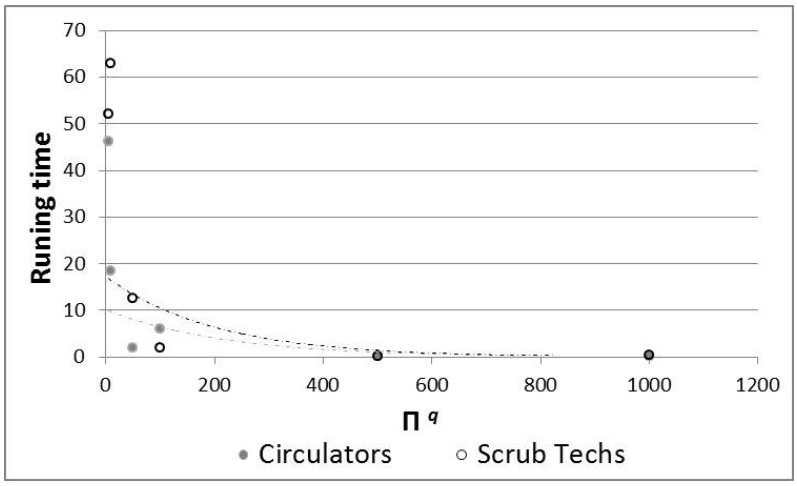


Figure 14: Circulators and scrub techs running time (minutes) vs. penalty Π^q for deviating from current staffing structure.

Table 9: Ratio of the heuristic's and optimal *Penalty* values, and the heuristic running time (minutes).

Staff Type	Budget (B_s)	Max. Different Shifts (N_J^{max})	$\Pi^q = 1000$		$\Pi^q = 500$		$\Pi^q = 100$	
			Ratio	Time	Ratio	Time	Ratio	Time
Circulators	CP	CP	1.10	6.6	1.10	6.3	1.03	13.5
		Flexible	1.27	2.6	1.27	5.9	1.06	6.34
	Phase I	CP	1.31	11.3	1.31	12.4	1.02	14.1
		Flexible	1.32	1.9	1.19	3.4	1.03	12.4
Scrub Techs	CP	CP	1.10	20.2	1.05	20.5	1.00	23.7
		Flexible	1.06	0.7	1.06	0.9	1.01	1.55
	Phase I	CP	1.02	10.4	1.02	11.9	1.00	15.7
		Flexible	1.03	1.0	1.03	1.7	1.03	3.6

larger than the optimal. With $\Pi^q = 100$ the results for *Penalty* are closer to the optimal, with a difference of 6% or less (less than 3% in most cases), and the running time for the heuristic is less than 24 minutes in all instances.

We define $Gap^H\%$ and $Gap_s^H\%$ as the measure of the gap between the required staffing levels and those resulting from the heuristic's staffing structure, similar to equations (31) and (32). Then, we define $\Delta Gap\%$ and $\Delta \sum Gap_s\%$ as follows:

$$\Delta Gap\% = Gap^H\% - Gap\% \tag{33}$$

$$\Delta \sum Gap_s\% = \sum_{s \in S} Gap_s^H\% - \sum_{s \in S} Gap_s\% \tag{34}$$

In Table 10, we report $\Delta Gap\%$ and $\Delta \sum Gap_s\%$ with the objective of comparing the performance of the heuristic and Phase II solutions, with respect to the actual staffing levels observed during the planning horizon. The results for $\Delta Gap\%$ show that the performance of the Phase II original formulation and the heuristic solutions are very similar (for all settings of Π^q), with a difference of less than 0.5%. Also $\Delta \sum Gap_s\%$ is 2.5% or less (for $\Pi^q = 100$). The (original) Phase II formulation can be used when the number of different shifts N_J^{max} is large, or when the manager requires a staffing structure that is very similar to $Y_{s,j,d}^0$ with a high Π^q; and the heuristic can be used in other settings that take longer time to solve optimally.

Table 10: $\Delta Gap\%$ and $\Delta \sum Gap_s\%$ for the heuristic's solutions under different penalty Π^q.

Staff Type	Budget (B_s)	Max. Diff. Shifts (N_J^{max})	$\Pi^q = 1000$		$\Pi^q = 500$		$\Pi^q = 100$	
			$\Delta Gap\%$	$\Delta \sum Gap_s\%$	$\Delta Gap\%$	$\Delta \sum Gap_s\%$	$\Delta Gap\%$	$\Delta \sum Gap_s\%$
Circulators	CP	CP	0.01%	1.08%	0.01%	1.08%	-0.01%	0.52%
		Flexible	0.05%	2.28%	0.05%	2.45%	-0.02%	0.29%
	Phase I	CP	0.40%	5.69%	0.40%	5.69%	0.33%	2.51%
		Flexible	-0.07%	3.39%	-0.12%	2.58%	-0.14%	1.18%
Scrub Techs	CP	CP	0.22%	1.61%	0.02%	0.73%	-0.06%	-0.15%
		Flexible	0.08%	1.05%	0.08%	1.05%	0.02%	0.33%
	Phase I	CP	0.00%	0.23%	0.00%	0.27%	0.01%	0.00%
		Flexible	0.07%	0.32%	0.04%	0.37%	0.04%	0.37%

2.4.5 Evaluation of Results by Simulation

With the Phase II formulation, we focus on measuring the difference between the budgeted staffing levels and the actual staffing level requirements. To directly evaluate our results under relevant operational measures such as the percentage of cases using 'pooled' staff (i.e., staff from other service lines), and the percentage of cases with delays, we develop a simulation model of the OR operations.

A general diagram of the OR patient flow is depicted in Figure 15. First, the patient arrives to the waiting area and some paperwork is done. We separate the patient flow into inpatients and outpatients, each with corresponding distributions for their arrival time relative to their scheduled procedure time. Once the procedure scheduled time is within a time window (around 2 hours), the patient is brought to the preparation area, where in some cases blood is drawn (more often in the case of outpatients), and then a nurse assesses and prepares the patient for the surgical procedure. This preparation time also depends on the type of patient (inpatient or outpatient). Once the patient, the OR, and the staff are ready for the case, the patient is brought to the OR. After the surgical procedure is done, the patient is taken to the recovery area, and the OR is set up for the next case by the OR staff, i.e., the OR is 'turned over' for the next case. The details of the simulation model implementation in Arena and its validation are in Appendix C.

We simulate the OR operations using Phase II solutions under each of the four

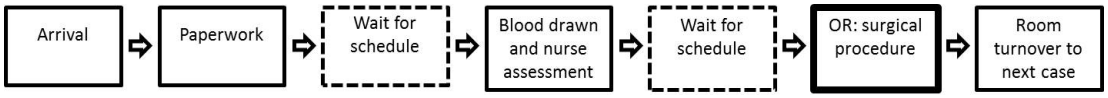


Figure 15: Concept ORs patient flow diagram of the simulation model.

Table 11: Simulation results under the hospital's current practice (CP) and Phase II staffing structures, for the percentage of cases delayed, the percentage of cases delayed because of OR staff unavailability, average OR staff delay (minutes), and the percentage of cases with staff from other service lines (staff pooling).

Budget (B_s)	Max. Diff. Shifts (N_J^{max})	Delayed Cases	Delayed Cases: OR Staff	Avg. OR Staff Delay	Circulators Pooling	Scrub Techs Pooling
CP staffing structure		*42%*	*7%*	*9.7*	*22%*	*9%*
CP	CP	39%	3%	0.93	18%	7%
	Flexible	40%	4%	2.27	18%	6%
Phase I	CP	40%	5%	2.26	18%	6%
	Flexible	39%	4%	1.64	18%	6%

settings in Table 6, and compare these simulation results with those obtained using the hospital's CP staffing structure. According to the simulation results (see Table 11), the overall percentage of delayed cases is reduced from 42% to 39-40% (depending on the Phase II setting). If we consider the percentage of delayed cases for which the unavailability of OR staff is the cause of delay, the percentage of delay is reduced from 7% to 3-5%. The largest improvement is on the average time of the delays caused by the OR staff unavailability with a reduction of 7.43-8.8 minutes with respect to the hospital's CP average time (9.7 minutes). There is also a reduction in the percentage of cases that took staff from other service lines, i.e., staff pooling, with a larger improvement in the case of the scrub techs. This is consistent with the results shown in Table 5, where the major reductions in (weekly) staff pooling are observed in scrub techs.

2.5 Conclusions

We consider the staff planning problem for a surgical department and propose a two-phase planning approach. In Phase I, we decide on the number of permanent FTEs to budget for each staff type, where staff is assigned to a particular service line and

each service line covers a set of surgical services. However, staff can be pooled (i.e., taken from other service lines) when enough staff of the corresponding service line are not available. We propose a model that uses historical surgical data to forecast the expected weekly staff hours demand during the planning horizon, considering growth trends and other factors affecting the demand, and generate different demand scenarios. Using these demand scenarios, we find the number of permanent FTEs that minimizes the expected labor costs, subject to limitations on the amount of pooled labor and use of overtime. We test the results using the actual surgical data for the planning horizon and are able to reduce weekly staff pooling by reassigning staff among the service lines without increasing the staff budget. Also, we find that only a small percentage of pooling and overtime are needed to reduce the recommended number of permanent FTEs.

In Phase II, we decide how the number of FTEs determined in Phase I should be allocated across the potential shifts and days, i.e., we determine the staffing structure. Phase II minimizes the penalty given the difference (gap) between the required staffing levels estimated by historical surgical data and the available staff according the staffing structure. The formulation considers restrictions in the staff planning and scheduling process such as the maximum number of shift assignments per week and per day, for a staff member. It incorporates constraints that make the staffing structure implementation easier. For instance, the OR manager may want to limit the number of different shifts or reduce changes to the current staffing structure. We observe that the Phase II running time increases after adding constraints to limit the number of different shifts. By contrast, considering a current staffing structure solution and a penalty for deviating from it decreases the running time. We incorporate these two observations in a heuristic that achieves almost the same results as the original formulation, in a fraction of the time. We test Phase II results with the actual surgical data for the planning horizon and are able to reduce the gaps

between the required and available staffing levels, without increasing the number of FTEs. Finally, we also test Phase II solutions with a simulation model of the OR operations and are able to reduce the number of cases with start time delays given insufficient OR staff, the average length of the delay because of OR staff, and the cases with staff 'pooled' from other service lines. We also develop a decision-support tool, described in Appendix D, to analyze demand trends and potential changes to the current staffing structure. In the future, we plan to explore the integration of Phase I and Phase II to this tool.

Phase II is a shift scheduling model and does not explicitly produce tours, which can directly model and warranty conditions that may be required such as continuous days-off. However, Phase II gives a lot of flexibility in terms of modeling a great variety of shifts with different start times and lengths, and service lines inclusive, which can become computationally intractable if we are modeling tours. The hospital of our case study already works with a shift schedule, and the OR coordinators try to satisfy most of the individual requests when assigning tours to the OR staff. Nonetheless, shifts schedules are commonly used in ORs [53, 45, 46, 116], and we believe they can also be useful in other similar healthcare settings were demand is mostly scheduled during the day on weekdays such as interventional radiology and some types of specialty clinics. In fact, Morris et al. [96] find that simpler shift scheduling models may work almost as well as more complex tour scheduling models (deeply studied in the literature [57]) under this type of demand settings; however, the authors stress that this 'myopic' approach is not adequate when the setting progresses more to a more continuous operating environment where more integrated scheduling approaches might be needed. This limits the applicability of the shift scheduling models, like the one we propose, to other common healthcare settings like acute care wards.

In our proposed solution approach, the decisions are made in a hierarchical way, as

it is currently done in practice: First the number of permanent FTEs is defined (the staffing budget), and secondly these FTEs are allocated to shifts and days (the staffing structure), which are later bundled and assigned to staff during the staff scheduling. Future work could explore the integration of these decisions in one model and explore if a more integrated - and potentially more complex - model could substantially increase the quality of the solutions in this type of settings.

CHAPTER III

WORKFORCE MANAGEMENT AND SCHEDULING UNDER FLEXIBLE DEMAND

3.1 Introduction

Labor is a major cost component in many industries. Salaries alone can account for more than 50% of all the operating expenses in service industries such as healthcare (52%), for-profit services (50%), and education services (50%) [110]; and salaries account for about 67% of total expenses in the service desk industry [93]. Hence, workforce planning and scheduling is crucial, since an improvement in labor productivity and staff satisfaction can translate into significant savings and reduce staff turnover.

This work is motivated by the staff planning and scheduling decisions of a company that offers document processing and other back-office services to healthcare providers. The company needs a schedule that focuses on on-time demand fulfillment but also considers staff preferences and operational practices. The document processing industry provides digital imaging, data capture, and analysis services to industries such as healthcare, insurance, finance, legal, government, etc., where processed documents include healthcare or insurance claims, credit card applications, and other forms and reports. These documents arrive by mail or electronically, and there is almost always a strict deadline by which the document processing must be completed. Hence, each arriving document should be processed within a time window, which requires not only a appropriate planning and scheduling of the staff, but also making decisions as to when each document should be processed (i.e., demand scheduling). Additional relevant characteristics of this problem include (i) capacity constraints regarding the

number of workstations or space available, which limit the maximum number of staff that can be on duty at any given time; (ii) employee differences in productivity and skill levels; and (iii) particular preferences or availability regarding work days and shifts. Other relevant applications can be found in warehouses or fulfillment centers, where the customer places an order (online or by phone), and the center's staff needs to pick and ship the order within a time window based on a promised service level. Similar to back-office services, the availability of equipment or workstations as well as staff preferences and characteristics impact the scheduling decisions.

We propose a mathematical model for simultaneous staff and demand (order) scheduling, given a time window for on-time demand fulfillment, while considering staff characteristics and preferences. The model is general and can be applied to many service settings. We implement our model and proposed solution methods for a healthcare back-office services provider, in the presence of additional implementation requirements of this company, such as team leader scheduling and minimum demand coverage requirements. The objective is to find a staff schedule that (i) can fulfill the demand on time, and (ii) it is a high-quality schedule in terms of the staff preferences and utilization. In Section 3.2 we review the relevant literature, and in Section 3.3 we describe the model. We discuss the model implementation (Section 3.4), additional constraints in our case study (Section 3.4.1), numerical experiments to develop insights on the model and its parametrization (Section 3.4.2), and the performance of a proposed heuristic (Section 3.4.3). Conclusions and future research directions are given in Section 3.5.

3.2 Literature Review

There is an abundant literature on staff planning and scheduling. Ernst et al. [58] give an annotated bibliography of about 700 references ranging from 1954 to 2004. Literature surveys include [5, 22, 57, 113] and [114]. Several of these surveys identify

the integration of decisions around demand and staff planning and scheduling as a potential research direction, which is the focus of our work.

Following the staff demand classification paradigm by Ernst et al. [57], our problem setting is related to the task-based demand (also called timetable demand) staff scheduling problem, where staff is scheduled and demand is assigned to staff. Here, the demand for staff consists of individual tasks, each defined by an earliest start time, latest completion time, and a duration. A typical application of this problem when tasks have different locations is vehicle crew scheduling, which requires defining a feasible sequence or route of tasks, followed by the staff assignment to these routes (see [7] and [82] for surveys on airline crew pairing and crew rostering, respectively). In many cases, the staff skills required to complete different tasks need to be taken into account. Firat and Hurkens [61] consider staff skills and availability to schedule multi-skilled tasks, but they do not consider different shifts (only full workdays). Begur et al. [23] propose a method to create routes for home-care nurses based on the daily demand as well as the nurses' availability and skills, with the objective of minimizing travel time (as a proxy to maximizing nurses' productivity). Given the complexity of this problem, few of the vehicle crew scheduling papers integrate both demand and staff scheduling (e.g., [62, 70, 79, 92]).

Integrating demand and staff scheduling decisions can reduce costs significantly [6]. Nevertheless, demand and staff scheduling decisions are often made sequentially in practice, and staff levels are either a result of the production schedule or considered as resource constraints [6, 9, 69]. Hanssmann and Hess [71] consider combining production and staffing decisions, which are limited to determining the workforce size. Other models that aim to integrate job-shop scheduling with staff planning and scheduling often ignore staff shifts and assume that staff is always available [1, 43, 59, 84], while in practice facilities often operate during more than one shift per day. Artigues et al. [9] and Guyon et al. [69] consider three non-overlapping shifts with the same duration

(staff can be assigned to only one of each three consecutive shifts to allow time for rest), and minimize the jop-shop makespan and the cost of the staff schedule. Our problem is closer to the one presented by Guyon et al. [68], where a set of tasks should be completed by the available staff within a time window. A set of potential schedule rosters (i.e., a bundle of shifts that can be assigned to one worker) is established for each staff member (according to individual preferences, contractual restrictions, etc.) along with an assignment cost. This implies that all the potential shift rosters or *tours* for each staff member should be considered, which may not be convenient when the workforce and the number of potential rosters (considering different shifts and days-off patterns) can be very large.

In traditional task-based demand applications, each task is modeled independently as a discrete entity. The applications we focus on are more-closely related to the setting proposed by Berman et al. [24], inspired by the operations of the U.S. postal service, where demand flows at high rates, such that employees can process many items per unit of time. There is also a time window for the arriving demand to be completed. They consider a network of workstations, but given the high volume, the flows are approximated as deterministic and the workforce scheduling problem is solved using a linear program. Bard et al. [18] present a stochastic demand model for the U.S. postal service staff scheduling problem; however, the (stochastic) demand has to be fulfilled at the time of arrival. Both papers propose methods to define staff requirements by shift but not by individual staff and demand fulfillment schedules. In fact, van den Bergh et al. [114] mention productivity (a relevant staff characteristic in our setting) as an example of one of the least-frequent personnel characteristics considered in staff scheduling models. Our work fills this void in the literature by considering and scheduling individual staff members.

Our model addresses the need to integrate the different modules of the staff planning and scheduling problem: demand and workforce planning, shift scheduling, shift

rostering, and roster assignments to staff. Moreover, the model is flexible as it (i) considers a set of penalties to balance both the demand fulfillment and the staff scheduling objectives, (ii) enables the customization of potential shifts and scheduling rules, (iii) considers several staff characteristics such as availability, preferences, and productivity, and (iv) allows the user to predefine decisions regarding the workforce composition or let the model make the decision.

Another contribution is the application itself. Most literature on staffing problems with applications in the service industry centers on operations where demand should be fulfilled upon arrival or soon after, such as call centers [40, 39]. Also, while most of the current literature on healthcare personnel scheduling focuses on nurses and other caregiver staff [57, 29], we focus on the back-office staff. A study by the Bank of New York Mellon Treasury Services [14] stated that more than half of healthcare transactions are still paper-based and manually processed, and healthcare providers spend over $100 billion to manage overall claims; hence, there is great potential for cost reduction. In addition, bad debt expenses average 12% to 13% of revenue, and the bills collection cycle averages more than 90 days [94], making efficiency and efficacy in back-office healthcare operations very important. Despite this, there is not much literature on back-office applications despite their relevance, particularly in industries such as healthcare. Huq et al. [75] propose a makespan minimization lot sizing model for a flow shop, with a practical application in document and payment processing. However, in their case, the daily volume is constant, and the workforce size and schedule are part of the model input. To the best of our knowledge, ours is the first work that fully addresses staff planning and scheduling in the document processing industry. Finally, we use real data and provide details on this research implementation (Appendices F and G), while in the current literature "... authors hardly ever provided details of the process of implementation or the observed results, although these could have been of interest to the reader" [114].

3.3 The Workforce and Demand Scheduling Model (WDSM)

We propose a Mixed Integer Program (MIP) to solve the workforce and demand scheduling problem. The model considers decisions about who among the available employees is scheduled in each time period during a planning horizon, which shifts to assign to each employee, and how to fulfill the forecasted demand considering employees' productivity and schedule, as well as the demand fulfillment time window. WDSM ensures that the selected shifts are feasible (i.e., allow enough rest and comply with the maximum number of scheduled hours per employee) and the capacity constraints (the maximum number of employees scheduled at a given time) are not violated. WDSM can schedule individual employees or clusters of employees and takes into account employee preferences and availability (individually or as a cluster).

WDSM's nomenclature is as follows:

Sets

$I = \{1, \ldots, N_I\}$	Employees (or clusters of employees)
$I^{fix} \subseteq I$	Employees who must be included in the schedule
$W = \{1, \ldots, N_W\}$	Weeks in the planning horizon
$D = \{-7, \ldots, -1, 1, \ldots, 7N_W\}$	Days, where $\{-7, \ldots, -1\}$ represent days in the previous schedule
$T = \{-7b - 1, 1, \ldots, 7bN_W\}$	Time buckets, where b is the number of time buckets per day and $\{-7b, \ldots, -2, -1\}$ represent time buckets in the previous schedule
$J = \{1, \ldots, N_J\}$	Shifts
$L = \{1, \ldots, N_L\}$	Locations
Δ	Set of ordered time pairs (t_1, t_2) that fall within the demand fulfillment time window, given a demand arrival at $t_1 \leq t_2$ and a fulfillment at t_2

V_t	Set of shift-day pairs (j, d) that cover time bucket t
Q^{stro} (Q^{soft})	Set of shift-day pair duples $((j_1, d_1), (j_2, d_2))$ that cannot (are not desired to) be assigned to the same employee
P_i	Set of shift-day pairs (j, d) where employee i was assigned to shift j on day $d = \{-7, \ldots, -1\}$
O_i	Set of shift-day pairs (j, d) where employee i is not available during shift j on day $d = \{1, \ldots, 7N_W\}$
$ESP = \{(i, j, d) : i \in I, (j, d) \notin O_i\}$	Set of all the feasible employee, shift, and day triples (i, j, d)

Parameters

F_t	Units of forecasted demand to arrive ($t \geq 1$) or past demand ($t \leq -1$) that was not fulfilled and should be completed during the planning horizon
δ	Time window (number of time buckets) allowed for demand fulfillment, i.e., demand arriving at time t must be fulfilled by time $t + \delta$
α	Fraction of demand that should be completed during the planning horizon, even if it could be delayed without violating the demand fulfillment time window
H_j	Length (hours) of shift j
E_i	Number of employees in cluster i. If modeling individual employees then $E_i = 1$
E_i^{prod}	Productivity of employee i (units of demand per time bucket)
$E_{i,w}^{h_{max}}(E_{i,w}^{h_{min}})$	Maximum (minimum) number of hours employee i can (should) be scheduled in week w, considering all shifts starting in week w
E_i^{loc}	Location of employee i
$M_{t,l}$	Maximum number of employees to schedule at time bucket t at location l

Penalties:

Π'	Penalty for each unit of demand not completed within the time window
Π_i	Penalty for scheduling employee i
$\Pi_{i,(j_1,d_1),(j_2,d_2)}^{soft}$	Penalty for scheduling shift-day pair duple $((j_1, d_1), (j_2, d_2)) \in Q^{soft}$ to employee i

$\Pi_i^{h_{min}}$	Penalty for scheduling fewer than $E_{i,w}^{h_{min}}$ hours to employee i
$\Pi_{i,j,d}^{pref}$	Penalty for assigning shift-day pair (j,d) to employee i
$\Pi_{i,w}^{dif}$	Penalty for assigning a shift to employee i in week w which is different than the shift assigned on the same day in the previous week
Π_i^{idl}	Penalty for idle time of employee i given the demand fulfillment schedule (measured as any additional units of demand that employee i could fulfill during a time bucket, considering E_i^{prod})

Variables

Z_i	1 if employee i is scheduled; 0, otherwise
$X_{i,j,d}$	1 if employee i is assigned to shift j, on day d, where $(i,j,d) \in ESP$; 0, otherwise
$S_{i,(j_1,d_1),(j_2,d_2)}$	1, if shift-day pair duple $((j_1,d_1),(j_2,d_2)) \in Q^{soft}$ is assigned to employee i, where $((j_1,d_1),(j_2,d_2)) \notin O_i$; 0, otherwise
$B_{i,w}$	Number of hours below $E_{i,w}^{h_{min}}$ for employee i during week w
$C_{i,j,d}^{+}(C_{i,j,d}^{-})$	1, if shift j was (was not) assigned to employee i on the same day the previous week and is not (is) assigned the current week, where $(i,j,d) \in ESP$, $d \geq 1$; 0, otherwise
$Y_{i,t}$	Units of demand scheduled to employee i at time bucket $t \geq 1$
$K_{i,t}$	Additional units of demand that employee i could fulfill during a time bucket $t \geq 1$ considering E_i^{prod} (i.e., a measurement of idle time)
$A_{(t_1,t_2)}$	Units of demand forecasted to arrive at t_1 and scheduled at t_2, $(t_1,t_2) \in \Delta$
F_t'	Unfulfilled units of (forecasted) demand at time $t \leq 7bN_W - \delta$
F''	Unfulfilled units of (forecasted) demand during time interval $[7bN_W - \delta < t \leq 7bN_W]$

We create a schedule for a predetermined time horizon consisting of N_W weeks. At the beginning of the planning horizon, the information P_i about the previous (rolling) schedule for each employee i is available to enable a feasible transition between schedules (e.g., allowing enough rest between the end of the previous schedule and the beginning of the new one). The set of potential shifts J is determined in advance. Each shift j is defined by a starting time and a duration H_j. Each day within the planning horizon is divided into b time buckets. For example, if the time window δ is 20 hours and a shift can last 8 or 12 hours, potentially starting every 4 hours after 7:00AM, then 4-hour time buckets (or six time buckets per day) starting at 7:00AM

would be reasonable. We consider two different sets of employees (or employee clusters): all employees $i \in I^{fix}$ must be scheduled, and the model decides whether or not to schedule employees $i \notin I^{fix}$ (e.g., temporal employees). Each employee or cluster of employees i has particular characteristics such as productivity E_i^{prod} and unavailable shifts-day pairs O_i (e.g., scheduled days-off, night shifts, etc.). The penalties $\Pi_{i,j,d}^{pref}$ can be set based on employee preferences regarding working certain days and shifts or based on the company preferences, e.g., to rotate employees through different days and shifts. Employees may be at different locations, and for each location $l \in L$ there might be a limit $M_{t,l}$ on the number of scheduled employees at a time t, e.g., given the available number of computers or workstations. There is also a limit on the maximum number of hours $E_{i,w}^{h_{max}}$ each employee i can be scheduled to work during week w, and a desired minimum of scheduled hours $E_{i,w}^{h_{min}}$. Given the company's scheduling rules (e.g., regarding the minimum rest time between two consecutive shifts), there are shift-day pair duples Q^{stro} that cannot be assigned to the same employee (e.g., a Monday-night shift followed by a Tuesday-morning shift). Also, there are shift-day pair duples Q^{soft} that could be assigned to the same employee but are not desired (e.g., a Monday-morning shift followed by a Tuesday-night). Moreover, it is sometimes desirable to have consistency in an employee's schedule during consecutive weeks (for instance, working consecutive Mondays in the morning), to facilitate employees' and managers' planning decisions such as employee transportation. This consistency is tracked with $C_{i,j,d}^{+}$ ($C_{i,j,d}^{-}$). Finally, the penalty for idle time Π_i^{idl} can vary by employee, e.g., a larger penalty for more-senior employees. Note that when modeling clusters of employees, we assume that the availability, scheduled hours requirements, and the location of the employees are homogeneous within the cluster, and that the productivity and penalties of the cluster is the sum of the cluster's employees' individual parameters.

WDSM balances the need to fulfill the forecasted demand in a timely fashion

with the desire to create an implementable schedule that considers each employee's (cluster's) characteristics and preferences. Since these objectives may conflict, we use a set of penalties in the objective function that can be adjusted so that the resulting schedule balances the different objectives.

Objective Function (OF)

$$MIN \quad Cost = \Pi' \Big[\sum_{\substack{t \in T: \\ t \leq 7bN_W - \delta}} F'_t + F'' \Big] \tag{35a}$$

$$+ \sum_{i \in I} \Pi_i Z_i \tag{35b}$$

$$+ \sum_{\substack{i \in I, ((j_1,d_1),(j_2,d_2)) \in Q^{soft}: \\ (j_1,d_1),(j_2,d_2) \notin O_i}} \Pi^{soft}_{i,(j_1,d_1),(j_2,d_2)} S_{i,(j_1,d_1),(j_2,d_2)} \tag{35c}$$

$$+ \sum_{i \in I, w \in W} \Pi_i^{h_{min}} B_{i,w} \tag{35d}$$

$$+ \sum_{i \in I, t \in T: t \geq 1} \Pi_i^{idl} K_{i,t} \tag{35e}$$

$$+ \sum_{\substack{(i,j,d) \in ESP: \\ d \geq 1}} \Pi^{pref}_{i,j,d} X_{i,j,d} \tag{35f}$$

$$+ \sum_{w \in W} \sum_{\substack{(i,j,d) \in ESP: \\ 7(w-1) < d \leq 7w}} \Pi^{dif}_{i,w} \big[C^+_{i,j,d} + C^-_{i,j,d} \big] \tag{35g}$$

OF minimizes the sum of penalties due to:

- Not fulfilling the forecasted demand within the time window (35a)
- Including an employee in the schedule (35b)
- Assigning an undesirable shift-day pair duple to an employee (35c)
- Not scheduling the minimum number of hours for an employee (35d)

- Employee idle time (35e)
- Assigning an undesirable shift-day pair to an employee (35f)
- Changing the weekly schedule for a given employee with respect of the previous week (35g)

Demand Scheduling Constraints (DSC)

$$\sum_{t_2 \in T:(t_1,t_2)\in\Delta} A_{(t_1,t_2)} + F'_{t_1} = F_{t_1} \quad t_1 \in T:\ t_1 \leq 7bN_W - \delta \tag{36}$$

$$\sum_{t_1 \in T:(t_1,t_2)\in\Delta} A_{(t_1,t_2)} = \sum_{i\in I} Y_{i,t_2} \quad t_2 \in T:\ t_2 \geq 1 \tag{37}$$

$$\sum_{\substack{(i,j,d)\in ESP:\\ (j,d)\in V_t}} E_i^{prod} X_{i,j,d} - K_{i,t} = Y_{i,t} \quad i \in I,\ t \in T:\ t \geq 1 \tag{38}$$

$$\sum_{\substack{t_1\in T:\\ t_1 > 7bN_W - \delta}} \sum_{t_2 \in T:(t_1,t_2)\in\Delta} A_{(t_1,t_2)} + F'' \geq \alpha \sum_{\substack{t_1\in T:\\ t_1 > 7bN_W - \delta}} F_{t_1} \tag{39}$$

Constraints (36) allocate the demand to time buckets within the time window and track the demand that is not scheduled. Constraints (37) allocate the demand to the employees' scheduled time buckets. Constraints (38) ensure that there is enough processing capacity during each scheduled time bucket to process this allocated demand, while measuring idle time as the $K_{i,t}$ units of demand not allocated to employee i during time t, given each employee schedule and productivity. Constraint (39) schedules at least a proportion α of the demand arriving during the end of the planning horizon (which could be feasibly scheduled later) and tracks the demand that is not scheduled within the planning horizon. Constraint (39) helps to transition to the next planning horizon.

Strong Scheduling Constraints (SSC)

$$Z_i = 1 \quad i \in I^{fix} \tag{40}$$

$$X_{i,j,d} = 1 \quad (j,d) \in P_i \tag{41}$$

$$X_{i,j,d} \leq Z_i \quad (i,j,d) \in ESP: \ d \geq 1 \tag{42}$$

$$\sum_{\substack{(i,j,d) \in ESP: \\ (j,d) \in V_t, E_i^{loc} = l}} E_i X_{i,j,d} \leq M_{t,l} \quad t \in T, \ l \in L: \ t \geq 1 \tag{43}$$

$$X_{i,j_1,d_1} + X_{i,j_2,d_2} \leq 1 \quad i \in I, \ ((j_1,d_1),(j_2,d_2)) \in Q^{stro}: \ (j_1,d_1), \ (j_2,d_2) \notin O_i \tag{44}$$

$$\sum_{\substack{(j,d) \notin O_i: \\ 7(w-1) < d \leq 7w,}} H_j X_{i,j,d} \leq E_i^{h_{max}} \quad i \in I, \ w \in W \tag{45}$$

SSC ensure that feasible schedules are created. Constraints (40) ensure that each fixed employee $i \in I^{fix}$ is scheduled. Constraints (41) ensure that shifts assigned to employee i in the previous schedule P_i are considered during the transition to the new schedule. Constraints (42) ensure that a shift is assigned to an employee only if he/she is scheduled. Constraints (43) ensure that the number of scheduled employees does not exceed the location's capacity (considering the employee clusters sizes). Constraints (44) ensure that a forbidden duple of shift-day pairs is not assigned to one employee, for instance, if there is not enough rest between the shifts. Constraints (45) limit the number of hours per week an employee is scheduled.

Flexible Scheduling Constraints (FSC)

$$\sum_{\substack{(j,d) \notin O_i: \\ 7(w-1) < d \leq 7w,}} H_j X_{i,j,d} + B_{i,w} \geq E_i^{h_{min}} Z_i \quad i \in I, \ w \in W \tag{46}$$

$$X_{i,j_1,d_1} + X_{i,j_2,d_2} - S_{i,(j_1,d_1),(j_2,d_2)} \leq 1$$
$$i \in I,\ ((j_1,d_1),(j_2,d_2)) \in Q^{soft} :\ (j_1,d_1),\ (j_2,d_2) \notin O_i \tag{47}$$

$$X_{i,j,d} + C^{+}_{i,j,d} - C^{-}_{i,j,d} = X_{i,j,d-7} \quad (i,j,d) \in ESP :\ d \geq 1,\ (j,d-7) \notin O_i \tag{48}$$

FSC help create 'desirable' schedules from the employees' and company's perspectives. Constraints (46) track the number of hours below the desired minimum per week. Constraints (47) keep track of assigning undesirable shift-day pairs duples to an employee. Constraints (48) keep track of the consistency of the schedule between consecutive weeks (by comparing a week's schedule to the previous one). Constraints (49)--(58) restrict the variables to be binary or non-negative.

$$Z_i \in \{0,1\} \quad i \in I \tag{49}$$

$$X_{i,j,d} \in \{0,1\} \quad (i,j,d) \in ESP \tag{50}$$

$$S_{i,(j_1,d_1),(j_2,d_2)} \in \{0,1\} \quad i \in I,\ ((j_1,d_1),(j_2,d_2)) \in Q^{soft} \tag{51}$$

$$B_{i,w} \geq 0 \quad i \in I,\ w \in W \tag{52}$$

$$C^{+}_{i,j,d}, C^{-}_{i,j,d} \in \{0,1\} \quad (i,j,d) \in ESP :\ d \geq 1 \tag{53}$$

$$Y_{i,t} \geq 0 \quad i \in I, t \in T :\ t \geq 1 \tag{54}$$

$$K_{i,t} \geq 0 \quad i \in I, t \in T :\ t \geq 1 \tag{55}$$

$$A_{(t_1,t_2)} \geq 0 \quad (t_1,t_2) \in \Delta \tag{56}$$

$$F'_t \geq 0 \quad t \in T :\ t \leq 7bN_W - \delta \tag{57}$$

$$F'' \geq 0 \tag{58}$$

3.3.1 The Fixed Schedule Model (FSM)

Consider a version of WDSM where the scheduled staff and their shifts are predetermined (i.e., Z_i and $X_{i,j,d}$ are given); then the only decisions are to schedule the forecasted demand given the available processing capacity (i.e, $Y_{i,t}$, $A_{(t_1,t_2)}$, F'_t, and F''). We call this the Fixed Schedule Model. FSM is relevant since it enables us to re-evaluate a particular staff schedule in case of a change in the model's input (e.g., forecasted demand or the length of the demand fulfillment time window) or a required adjustment in the schedule. In Section 3.4.3 we solve FSM as part of a proposed heuristic, and in Section 3.4.2.3 we use FSM to evaluate the generated schedule' robustness.

Proposition 3.3.1. *FSM is pseudo-polynomially solvable.*

To prove Proposition 3.3.1 (see Appendix E), we model FSM as a minimum cost circulation problem. The size of the network depends on the number of time buckets $|T|$.

3.4 Case Study

This research was motivated by a company that provides back-office services, such as chart workflow redesign, remote coding services, remittance processing, and denials workflow, to healthcare providers. Our focus is on the remote coding services, in particular, the operations (which employ more than 700 staff) related to the processing of explanation of benefits forms (EOBs). The process of an EOB starts with the generation of claim data by the healthcare provider (such as a hospital or a clinic), followed by the payer's approval of the claim and subsequent payment and generation of the EOB, the lockbox payment deposit and scan of the EOB, and finally the process of the EOB to generate data for the healthcare provider (see Figure 16). The demand, i.e, the EOB documentation, arrives after the payment is done, and the company has

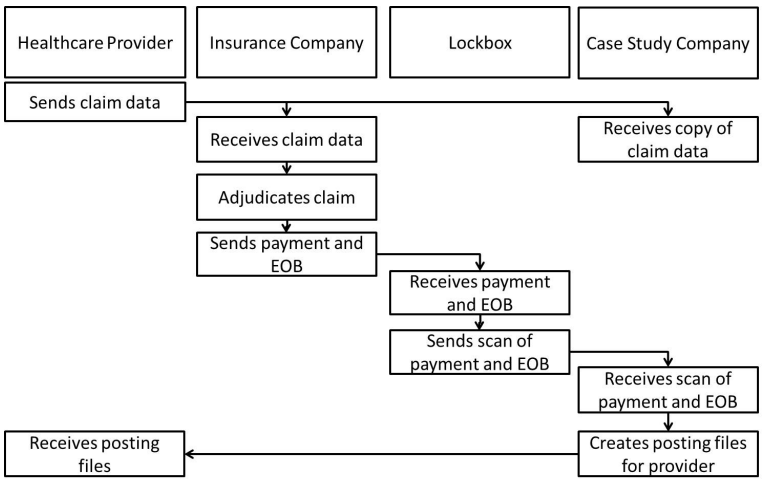


Figure 16: Explanation of benefits form (EOB) workflow.

a time window to generate the corresponding posting files. The company receives a forecast for the number of EOBs from its clients, which allows them to plan for their workforce in advance.

The company's goal was to improve the efficiency of their operations by obtaining a better match between the demand and the staff processing capacity, and as a result (i) improve their service level (i.e., on-time processing of claims), and (ii) reduce staff's idle time. Each employee is paid a small base salary and receives a reimbursement based on the amount of work he/she completes. Hence, staff would earn higher income with less idle time. Lower idle time combined with the consideration of staff's scheduling preferences would also reduce staff turnover in the long-term, leading to more experienced (and productive) staff.

We consider the following decisions:

1. Who is scheduled? (Workforce planning)

2. Which shifts are assigned to each scheduled employee? (Staff scheduling)

3. How should demand be allocated across the assigned shifts? (Demand scheduling and allocation)

WDSM (described in Section 3.3) models these decisions. The input parameters in WDSM's implementation are discussed in Appendix F. Details of the implementation and decision-support tools are presented in Appendix G.

3.4.1 Extensions to WDSM

The following additional constraints are added to WDSM to support the operational practices of the company: (i) team supervision, and (ii) a minimum coverage for specific training and skills. While it is generally desirable to maintain the same schedule for the entire team, in some cases an employee may not be scheduled with his/her leader due to different availability, location capacity restrictions, or the fact that breaking a team will improve demand coverage and staff utilization. The company's practice is to schedule a minimum number of leaders (at each location) and employees trained to serve each group of clients, such that there is always leadership and a minimum level of processing capacity for each client group at any time. The corresponding additional nomenclature and constraints are as follows:

Additional Sets

$R = \{1, \ldots, N_R\}$	Set of team leaders, $R \subseteq I$
I_r	Set of employees supervised by team leader r, $I_r \subseteq I$
$G = \{1, \ldots, N_G\}$	Set of client groups requiring specific type of employees
EG	Set of pairs (i, g) where employee i is eligible to serve client group g

Additional Parameters

$G_{g,t}^{min_p}$	Minimum scheduled productivity (units of demand/time bucket) at time $t \geq 1$ for group g
$G_{g,t}^{min_r}$	Minimum number of scheduled leaders at time $t \geq 1$ for group g at any location
E_i^{nr}	Number of leaders $r \in R$ in employee cluster i. Note that if $E_i = 1$ and $i \notin R$, then $E_i^{nr} = 0$

Penalties:

Π_i^{supr}	Penalty of scheduling employee i without his team leader for supervision
$\Pi_g^{min_p}$	Penalty for not achieving the minimum productivity for group g
$\Pi_g^{min_r}$	Penalty for not achieving the minimum number of team leaders for group g

Additional Variables

$S_{i,j,d}^{supr}$	1 if employee i is scheduled without his team leader r during shift j on day $d \geq 1$, where $i \in I_r$, $(j,d) \notin O_r \cup O_i$; 0, otherwise
$B_{g,t}^{min_p}$	Productivity gap below minimum at time $t \geq 1$ for client group g
$B_{g,t,l}^{min_r}$	Leadership gap below minimum at time $t \geq 1$ and location l for client group g

Additional Constraints (AC)

$$X_{i,j,d} - S_{i,j,d}^{supr} \leq X_{r,j,d} \quad i \in I_r,\ r \in R,\ j \in J,\ d \in D:\ d \geq 1,\ (j,d) \notin O_r \cup O_i \tag{59}$$

$$\sum_{\substack{(i,j,d) \in ESP: \\ (j,d) \in V_t, (i,g) \in EG}} E_i^{prod} X_{i,j,d} + B_{g,t}^{min_p} \geq G_{g,t}^{min_p} \quad g \in G,\ t \in T:\ t \geq 1 \tag{60}$$

$$\sum_{\substack{(i,j,d) \in ESP: \\ (j,d) \in V_t, (i,g) \in EG, \\ E_i^{loc} = l}} E_i^{nr} X_{i,j,d} + B_{g,t,l}^{min_r} \geq G_{g,t}^{min_r} \quad g \in G,\ t \in T,\ l \in L:\ t \geq 1 \tag{61}$$

$$S_{i,j,d}^{supr} \in \{0,1\} \quad i \in I_r,\ r \in R,\ j \in J,\ d \in D:\ d \geq 1,\ (j,d) \notin O_r \cup O_i \tag{62}$$

$$B_{g,t}^{min_p} \geq 0 \quad g \in G,\ t \in T:\ t \geq 1 \tag{63}$$

$$B_{g,t,l}^{min_r} \geq 0 \quad g \in G,\ t \in T,\ l \in L:\ t \geq 1 \tag{64}$$

Constraints (59) track the shifts without supervision by comparing the schedule of each employee with his/her leader's (if any). Constraints (60) and (61) keep track of the minimum demand coverage requirements for each client group. Constraints

(60) consider the processing capacity (productivity) of the workers trained to process demand from each group. Constraints (61) keep track of the leadership available by client group and location, in each time period. Finally, constraints (62)--(64) restrict variables to be binary or non-negative.

Case Study Objective Function (OF^{CS})

$$MIN \quad Cost^{CS} = MIN \quad Cost \quad + \sum_{\substack{i \in I_r, r \in R, j \in J, d \in D: \\ d \geq 1, (j,d) \notin O_r, O_i}} \Pi_i^{supr} S_{i,j,d}^{supr} \tag{65a}$$

$$+ \sum_{\substack{g \in G, t \in T: \\ t \geq 1}} \Pi_g^{min_p} B_{g,t}^{min_p} \tag{65b}$$

$$+ \sum_{\substack{g \in G, t \in T, \\ l \in L: t \geq 1}} \Pi_g^{min_r} B_{g,t,l}^{min_r} \tag{65c}$$

The objective function (65) minimizes the sum of penalties, including those specific to this case study due to not having the required leadership/supervision during an assigned shift (65a) and not scheduling the minimum productivity and leadership coverage for each group client (65b, 65c).

3.4.2 Computational Study

We run a series of experiments to gain insights on the effects of different input and operational settings on the resulting schedule and on-time demand fulfillment. In particular we are interested in answering the following questions:

1. What is an appropriate level for the unfulfilled demand penalty, Π'?
2. What are the effects of:
   (a) Changing the service terms, in particular, increasing or reducing the allowed time window for demand fulfillment?

(b) Changing the demand arrival behavior, in particular, if the demand arrives in smaller or larger batches (i.e., demand batching)? This was identified as a potential area of future flexibility by the company's management.

(c) Employee and company schedule preferences or 'soft' constraints?

(d) Company's practices described in Section 3.4.1 such as team leader supervision and minimum demand coverage for client groups?

3. How *robust* is WDSM's staff schedule under different demand scenarios (i.e, can the staff schedule handle uncertainties in demand without significantly affecting demand fulfillment)?

To address the first two questions, we test different levels for the following input settings (highlighted in **bold** is the baseline setting, which was eventually implemented in practice):

- Penalty for unfulfilled demand (Π'): 1, 5, **50**, 500, 5,000, 50,000
- Demand fulfillment time window (δ): 16 hours (4 time buckets), **20** hours (5 time buckets), and 24 hours (6 time buckets)
- Forecasted demand (F_t) batching: high, **medium**, and low (see Section 3.4.2.2 for details).

To address questions 2(c) and 2(d), we consider the following WDSM settings:

Setting	Description	Constraints	Objective Function
I	Creates a 'feasible' schedule but does not consider preferences and soft restrictions	DSC, SSC	35a, 35b
II	Considers preferences and soft restrictions	DSC, SSC, FSC	35a--35g
III	Considers preferences, soft restricttions, and additional constraints such as team leader supervision (specific to the case study)	DSC, SSC, FSC, AC	35a--35g, 65a--65c

We run each WDSM setting (I, II, and III) at each of the six proposed levels for the unfulfilled demand penalty Π' at the baseline time window ($\delta = 20$) and demand batching (i.e., 18 experimental settings), and at $\Pi' = 50$ and each of the additional eight combinations of demand batching and time window levels (i.e., 24 additional experimental settings). Hence, there are a total of 42 experimental settings (14 for each of the three WDSM settings).

Finally, to address the last question, we add a random deviation to the original demand forecast F_t (in setting III, with baseline parameters). We generate five random demand scenarios at three deviation levels (low, medium, and high) for a total of 15 experimental settings, and solve FSM under the original WSDM staff schedule and the alternative demand scenarios (see Section 3.4.2.3 for details).

Throughout the computational studies, we consider six 4-hour time buckets per day (starting at 7:00AM), and three 8-hour shifts: 7:00AM to 3:00PM, 3:00PM to 11:00PM, and 11:00PM to 7:00AM. We use WSDM and the company data (demand forecast, staff characteristics and availability, previous schedule, and preferences and penalties) for the last nine two-week periods in 2012 (see Appendix F for details). Each experimental setting is solved for each of these nine two-week planning periods using CPLEX. We set a bound on the optimality gap (i.e., the difference between the best known solution and a bound on the best possible solution) of 0.5%, and a running time limit of 24 hours.

Table 12: Results for WDSM Setting I.

Unfulfilled Demand Penalty	Time Window (hours)	Demand Batching	Average Unfulfilled Demand %	Average $\bar{\Pi}$	Average CPLEX Time (min.)	Average Optimality Gap %
1	**20**	**Medium**	1.2%	1.4	8	0.0%
5	**20**	**Medium**	1.0%	1.4	8	0.1%
50	16	Low	1.9%	1.4	18	0.1%
		Medium	2.2%	1.7	19	0.2%
		High	2.5%	1.8	14	0.2%
	20	Low	1.2%	1.4	15	0.1%
		Medium	0.8%	1.4	20	0.2%
		High	1.8%	1.7	19	0.2%
	24	Low	0.8%	1.4	13	0.1%
		Medium	0.7%	1.4	12	0.1%
		High	0.9%	1.4	23	0.2%
500	**20**	**Medium**	0.7%	1.4	15	0.2%
5,000	**20**	**Medium**	0.7%	1.4	19	0.1%
50,000	**20**	**Medium**	0.7%	1.4	24	0.1%

3.4.2.1 On-time Demand Fulfillment vs. Other Schedule Preferences

Tables 12, 13, and 14 show the results of WDSM settings I, II, and III, respectively, for each of the 14 input settings explained in Section 3.4.2, and each entry in the table represents the average of the results for nine two-week schedules. The percentage of unfulfilled demand is the percentage of the forecasted demand that could not be completed on time. To assess the schedule quality in terms of meeting preferences, we compute $\bar{\Pi}$, which is the preference penalty 'paid' per unit of demand fulfilled (i.e., the ratio of the penalties from (35c), (35e), (35f), and (35g) to the on-time fulfilled demand).

As expected, as the penalty for unfulfilled demand Π' increases, the percentage of unfulfilled demand decreases until no more demand can be completed (see Figure 17). While the penalty Π' does not significantly affect the percentage of unfulfilled demand in WDSM setting I (since this setting only considers the cost for scheduling staff, and ignores other preferences and soft restrictions), its effect is significant in WDSM settings II and III, where there is a trade-off between schedule preferences and demand fulfillment.

Table 13: Results for WDSM Setting II.

Unfulfilled Demand Penalty	Time Window (hours)	Demand Batching	Average Unfulfilled Demand %	Average $\bar{\Pi}$	Average CPLEX Time (min.)	Average Optimality Gap %
1	**20**	**Medium**	4.4%	8.3	2	0.0%
5	**20**	**Medium**	3.6%	8.2	2	0.0%
50	16	Low	2.2%	8.2	4	0.1%
		Medium	3.6%	9.0	5	0.1%
		High	4.3%	9.4	4	0.0%
	20	Low	1.4%	7.7	6	0.1%
		Medium	1.6%	8.3	7	0.0%
		High	2.9%	9.0	8	0.0%
	24	Low	1.0%	7.6	3	0.0%
		Medium	1.1%	8.0	4	0.1%
		High	1.4%	8.3	5	0.1%
500	**20**	**Medium**	0.8%	9.7	11	0.1%
5,000	**20**	**Medium**	0.7%	9.9	14	0.1%
50,000	**20**	**Medium**	0.7%	9.9	14	0.1%

Table 14: Results for WDSM Setting III.

Unfulfilled Demand Penalty	Time Window (hours)	Demand Batching	Average Unfulfilled Demand %	Average $\bar{\Pi}$	Average CPLEX Time (min.)	Average Optimality Gap %
1	**20**	**Medium**	5.8%	15.2	115	0.0%
5	**20**	**Medium**	5.3%	14.7	230	0.6%
50	16	Low	3.6%	14.8	106	0.4%
		Medium	5.4%	15.1	196	0.3%
		High	7.4%	16.2	389	0.6%
	20	Low	2.2%	14.1	105	0.0%
		Medium	3.0%	14.4	112	0.0%
		High	5.2%	14.9	137	0.7%
	24	Low	1.5%	13.3	61	0.1%
		Medium	1.9%	13.5	92	0.5%
		High	2.6%	14.6	148	0.4%
500	**20**	**Medium**	1.4%	16.5	168	0.7%
5,000	**20**	**Medium**	0.9%	18.7	308	0.6%
50,000	**20**	**Medium**	0.8%	19.4	245	0.4%

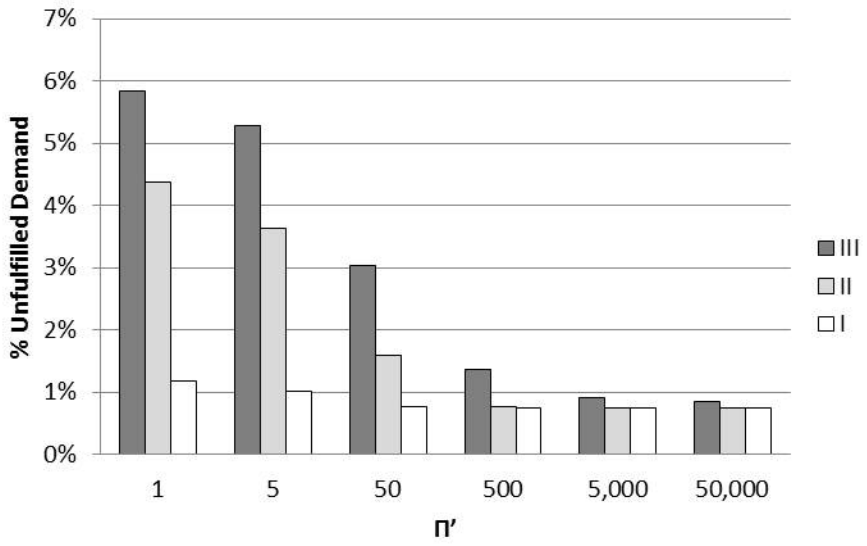


Figure 17: Average unfulfilled demand percentage vs. penalty Π' and WDSM settings I, II, and III.

Increasing the penalty for unfulfilled demand Π' increases the demand completed on time, but it could be at the expense of choosing a worse schedule in terms of meeting employee or company preferences. At WDSM setting III (with baseline parameters), $\bar{\Pi}$ (the preference penalty paid per unit of fulfilled demand) is lowest at $\Pi' = 50$. For this reason, we set $\Pi' = 50$ in the implementation to maintain a balance between the quality of the schedule and demand fulfillment.

3.4.2.2 Effect of Demand Time Window and Batching

The company would prefer the demand to arrive 'smoothly' over time and have a larger time window for fulfillment to reduce idle time. However, the company believes that their clients do a certain level of batching while submitting the EOBs for processing. To evaluate the effects of demand batching and the time window, we consider three levels of demand batching (i) low, where daily forecasted demand is uniformly distributed across the day's time buckets (less variability), (ii) medium (current practice), and (iii) high or 'lumpy', where daily forecasted demand arrives in two of the six time buckets during the day (more variability). To reflect potential

service terms agreements, we set three levels for the time window: 16 hours, 20 hours (current practice), and 24 hours.

Intuition suggests that as the demand variability increases, demand fulfillment would decrease. We observe this for WDSM settings II and III (see Tables 13 and 14) but not always for setting I (see Table 12). WDSM setting I has more flexibility in choosing among different schedules since schedule preferences and soft restrictions are not considered, and some demand batching may allow additional flexibility for demand fulfillment within a time window; for instance, time buckets of low demand allow staff to rest and be ready for upcoming time buckets of high demand (see Figure 44 in Appendix G for an example).

As the time window decreases, the unfulfilled demand increases (or remains the same) in all WDSM settings. The impact of a change in the time window is significant: for (the implementation) setting III, the average unfulfilled demand increases by more than 50% if the time window is reduced from 20 hours to 16 hours for the medium demand batching level (see Table 14).

3.4.2.3 Schedule Robustness Under Different Demand Scenarios

The case study company obtains a daily demand forecast from its clients and breaks it into time buckets based on historical data. The company then schedules its staff based on this demand forecast. We are interested in evaluating the robustness of WDSM's staff schedule under different demand scenarios. We start with the schedules for each of the nine two-week planning horizons, under the implementation baseline setting ($\Pi' = 50$, $\delta = 5$, and WDSM setting III) and the company's demand forecasts. We subtract/add a random percentage of the demand forecast in each time bucket, drawn from one of three uniform distributions: $U(-10\%, 10\%)$, $U(-20\%, 20\%)$, $U(-30\%, 30\%)$ (low, medium, and high, respectively), to generate

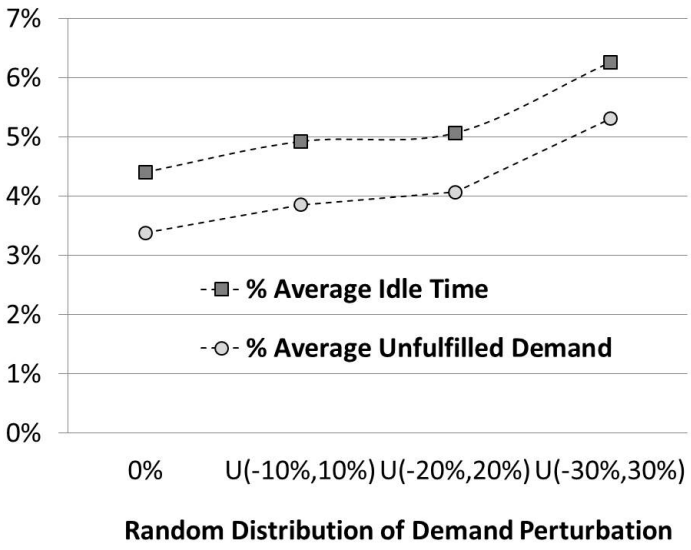


Figure 18: Average percentage of staff idle time and average percentage of unfulfilled demand, under different random perturbations of forecasted demand.

five random demand scenarios for each of the nine two-week schedules for each distribution (135 instances in total, 45 for each level of variability). We solve FSM after giving the new demand scenarios and WDSM's staff for the original demand forecast as input.

Figure 18 shows the results for the average percentage of staff idle time and the average percentage of unfulfilled demand. As the demand variability increases, the match between the scheduled capacity and the demand decreases (the average idle time and the average unfulfilled demand increase). Fortunately, even with an up to $\pm 20\%$ demand deviation (from forecast), there is only a slight increase for both the idle time and the average unfulfilled demand (less than 1%) compared with the *perfect information* scenario where demand equals the forecast. These results suggest that the staff schedules are robust when there is demand uncertainly.

3.4.3 Heuristic

WDSM addresses weekly or bi-weekly planning decisions; however, it is still important for the planners to obtain the model results quickly, since this would enable them to schedule their staff closer to the start of the scheduling horizon with more-accurate estimates of the input parameters, and to have the flexibility to run the model with different parameters and scenarios. Moreover, in this application, the commercial solver license runs in the cloud and it is billed for the time it is running; and hence a longer running time also impacts the scheduling costs. Tables 12, 13, and 14 give the average CPLEX running time for each experimental setting. WDSM settings I and II, especially the latter, run considerably faster than setting III. However, we are particularly interested in the implementation's baseline setting (WDSM setting III, $\Pi' = 50$, $\delta = 5$, and the original demand forecast), which has an average running time of approximately 2 hours, though some instances run 4 hours or longer. Figure 19 shows the average optimality gap versus runtime for this experimental setting. The optimality gap decreases significantly during the first two hours, but then the gap decreases slowly. We propose a heuristic with the goal of finding a feasible and good schedule quickly.

The general idea of the heuristic is to assign one shift at a time until completing the schedule for one employee, and then move to the next one. First, the employees are sorted such that the team leaders (if any) are scheduled first, followed by the rest of the staff in order of decreasing productivity. The heuristic starts assigning the first shift to the first employee on the list. It loops through employee i's feasible shift-day pairs (j, d), assigns available demand based on the shift time coverage (i.e., $t : (j, d) \in V_t$), the demand time window δ (starting with the time bucket t with more demand to complete), and the employee's productivity E_i^{prod}; and chooses the best shift-day pair (j, d) given the total penalty that results from this shift assignment based on the company and staff preferences and penalties. When an employee serves demand,

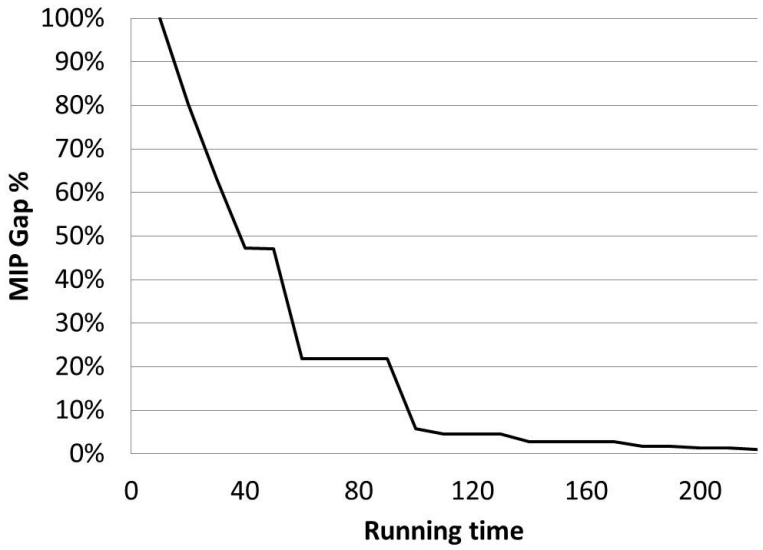


Figure 19: Average optimality gap in CPLEX vs. time (minutes) for WDSM setting III, $\Pi' = 50$, $\delta = 5$, and original demand forecast F_t.

there is an 'avoided' penalty due to the reduction of (35a). If the avoided penalty for serving this demand during this shift is not greater than the penalty resulting from assigning it (plus a portion of the penalty Π_i for including employee i in the schedule) times a factor defined by the planner, then the employee is not scheduled, i.e., $Z_i = 0$, unless he/she was originally 'fixed' to the schedule (i.e., $Z_i = 1$ if $i \in I^{fix}$). Otherwise, the first shift of the employee is assigned, i.e., $X_{i,j,d} = 1$, and the scheduled demand is allocated to the employee and the remaining demand forecast is updated. Then, the next shift of the employee is chosen until (i) there are no feasible shift-day pairs (j, d) to choose from (considering already assigned shifts), or (ii) the employee's minimum hours are already scheduled and the avoided penalty does not justify scheduling the additional shift. The heuristic then moves to the next employee. An outline of the heuristic's algorithm is shown in Appendix H.

This heuristic is 'myopic' because it decides one shift, one employee at a time, without considering the entire planning horizon and entire staff. To overcome this weakness, we introduce an *initial value* from zero to one, for each employee-shift-day

triple $(i, j, d) \in ESP$, $d \geq 1$, and an additional penalty, Π^{dev}, for deviating from this value. For instance, if $\Pi^{dev} = 1,000$ and the triple (i, j, d) has an initial value of 0.4, then there is an additional penalty of $1,000(1 - 0.4) = 600$ for assigning shift-day (j, d) to employee i, which is added to the previously computed total penalty for (j, d). Therefore, these initial values favor/unfavor particular shift-day assignments for a given employee. These initial values should be obtained quickly and provide a 'hint' of a good schedule that considers the entire planning horizon and staff. We consider three sources of initial values.

LP-based: Use the (fractional) solution of the linear relaxation of WDSM.

Cluster-based: Use the integer solution of WDSM with the employees clustered by teams, and those (few) employees without a team clustered by location.

LP-cluster-based: Use the (fractional) solution of the linear relaxation of the cluster-based WDSM above.

The initial values could also come from other sources, e.g., a previous schedule. Finally, we solve FSM to obtain the optimal demand fulfillment schedule subject to the the staff schedule produced by the heuristic.

3.4.3.1 Heuristic Performance

We test the proposed heuristic using the three sources for the initial values described in Section 3.4.3 and a high, low, and zero additional penalty Π^{dev}, under the baseline settings described in Section 3.4.2. The results are shown in Tables 15, 16, and 17 for WDSM settings I, II, and II, respectively; and each entry in a table represents the average of the nine two-week planning horizons (with the company data). We do not report the total running time since it is less than 3 minutes (including the time to obtain the initial values, construct a feasible schedule, and solve FSM, which solves almost instantaneously). To evaluate the quality of the schedules, we use the following

performance metrics (in addition to the unfulfilled demand, idle time percentages, and the preference penalty per unit of demand fulfilled $\bar{\Pi}$):

- Hours to schedule: Percentage of hours below the minimum required for each employee (and hence, subject to a penalty as expressed in (35d) with respect of the total number of hours (the scheduled hours plus those required to achieve the minimum). This metric is relevant as it affects staff reimbursement.

- Shifts without Leader: Percentage of shifts to which staff are assigned without their corresponding leader (penalty expressed in (65a)).

- Non-covered minimum productivity (leadership): Percentage of the minimum scheduled productivity (leadership), $G_{g,t}^{min_p}$ ($G_{g,t}^{min_r}$), that was not covered during the planning horizon for all client groups $g \in G$ (penalties expressed in (65b) and (65c), respectively).

Not all of these performance metrics apply to all of the model settings I, II, and III. For instance, $\bar{\Pi}$ is not relevant for setting I, which does not consider preferences.

For setting I, the heuristic with cluster-based initial values generates better schedules in terms of service level (100% − unfulfilled demand percentage) and idle time. This heuristic has averages of 2.9% unfulfilled demand and 0.4% idle time versus averages of 0.7% and 0.2%, respectively, in the optimal schedule (see Table 15, highlighted in *italic*). The results are not significantly influenced by the penalty for deviation from the initial values (Π^{dev}). The cluster-based initial values are integer and offer a good coverage of demand, and as preferences are not considered in this setting, the heuristic schedules those shifts as long as they are feasible. On the contrary, the LP-based initial values can be fractional, and the overall demand coverage can be lost by rounding up/down the initial values when scheduling an employee.

For setting II, the heuristic with LP-based initial values generates better schedules, in particular in terms of the metrics regarding scheduling preferences, which means

that they are desirable from the staff's and company's perspectives. The LP-based initial values come from the linear relaxation of WDSM, considering the staff preferences individually rather than as a cluster. In addition, they also consider the 'overall picture,' with better results in terms of demand fulfillment and capacity utilization than the heuristic setting with no initial values. Again, only a small additional penalty Π^{dev} is sufficient. The LP-based/low results are similar to the optimal with 2.8% average unfulfilled demand (versus 1.6%) and 6.8% average idle time (versus 3.3%), and a similar average $\bar{\Pi}$ as the optimal (see Table 16). The average percentage of hours to schedule does not vary significantly across the different heuristic settings.

For setting III, the heuristic with cluster-based initial values generates better schedules, particularly in terms of the 'coverage' performance metrics. The cluster-based initial values schedule the clusters so that they provide a good coverage of the demand by client group, assuming that the team members are available when their team leader is; and the heuristic makes adjustments in the schedules when this is not the case. Therefore, the heuristic does not need to adjust the team leader's schedule, resulting in a 17.5% average non-covered leadership, which is optimal (see Table 17). However, the heuristic with LP-based initial values performs better in terms of the percentage of shifts without the team leader, since the initial values already consider each team member's availability. Other performance metrics are also comparable with those of the optimal schedule: 3.9% average unfulfilled demand versus 3.0%, 6.8% average idle time versus 4.2%, and a 22.9 average $\bar{\Pi}$ vs. 19.13. Similar to the other model settings, only a small additional penalty Π^{dev} is sufficient. The near-optimal performance of the heuristic under setting III is the most relevant as (i) it is the implemented setting in this case study, and (ii) settings I and II solve relatively fast (see Tables 12 and 13).

Table 15: Heuristic Results for WDSM Setting I.

Initial Values Source	Initial Values Π^{dev}	Average Unfulfilled Demand %	Average Idle Time %
None		8.5%	1.4%
LP	Low	8.5%	0.5%
	High	8.6%	0.5%
Cluster	*Low*	*2.9%*	*0.4%*
	High	2.9%	0.4%
LP-cluster	Low	9.0%	0.5%
	High	9.6%	0.5%
Optimal Solution		*0.7%*	*0.2%*

Table 16: Heuristic Results for WDSM Setting II.

Initial Values Source	Initial Values Π^{dev}	Average Unfulfilled Demand %	Average Idle Time %	Average $\bar{\Pi}$	Hours to Schedule %
None		3.8%	8.4%	11.7	0.4%
LP	*Low*	*2.8%*	*6.8%*	*11.0*	*0.3%*
	High	2.7%	6.7%	12.6	0.3%
Cluster	Low	2.8%	7.2%	14.6	0.3%
	High	2.8%	7.3%	15.7	0.3%
LP-cluster	Low	4.3%	8.5%	14.7	0.4%
	High	4.9%	8.0%	17.2	0.4%
Optimal solution		*1.6%*	*3.3%*	*10.3*	*0.3%*

Table 17: Heuristic Results for WDSM Setting III.

Initial Values Source	Initial Values Π^{dev}	Average Unfulfilled Demand %	Average Idle Time %	Average $\bar{\Pi}$	Hours to Schedule %	Shifts Without Team Leader %	Non-covered Min. Productivity %	Non-covered Min. Leadership %
None		5.4%	8.9%	17.1	0.8%	3.3%	7.9%	34.6%
LP	Low	4.6%	7.3%	22.6	0.8%	1.9%	4.2%	24.6%
	High	4.3%	6.8%	23.0	0.9%	2.2%	3.9%	24.6%
Cluster	*Low*	*3.9%*	*6.8%*	*22.9*	*0.4%*	*3.8%*	*2.2%*	*17.5%*
	High	3.9%	6.8%	24.1	0.4%	3.7%	2.2%	17.5%
LP-cluster	Low	4.9%	7.4%	19.9	0.8%	3.6%	3.8%	25.4%
	High	5.2%	6.9%	21.3	0.9%	3.7%	3.6%	25.2%
Optimal solution		*3.0%*	*4.2%*	*19.13*	*0.3%*	*3.1%*	*1.6%*	*17.5%*

3.5 Conclusions

We considered a staff planning and scheduling problem where the demand needs to be fulfilled within a time window. We proposed an optimization model, which we implemented to schedule claim coders for a company that offers healthcare back-office services. It is an integrated model that considers the decisions regarding: workforce planning by deciding which employees to schedule, shift scheduling and shift roster assignments based on staff characteristics, preferences, and availability, and demand fulfillment scheduling based on a given forecast. We also implemented a set of decision-support tools to facilitate the model's input generation and output analysis. Using a computational study, we analyzed the trade-offs between the two main objectives: on-time demand fulfillment and the quality of the schedule (e.g., regarding staff utilization, adherence to staff preferences, etc.). We also analyzed the effects of two potential changes in the clients' behavior: demand batching and demand fulfillment time window.

We find that a small change in the demand fulfillment time window can have a significant effect on the demand fulfillment and the quality of the schedule, and that, in general, demand batching negatively affects performance metrics. We also analyze the effects of including the additional operational constraints of the company, which are presented in Section 3.4.1. These practices, such as team scheduling, could apply to other companies [41, 101]. We analyze the robustness of the model-generated staff schedule under different demand scenarios, and find that the percentages of unfulfilled demand and staff idle time are robust to errors in the demand forecast. Finally, we propose a heuristic to quickly construct staff schedules, which generates good schedules based on relevant performance metrics.

The company reported a 25% increase in staff productivity after WDSM's implementation (although other improvement projects were done in parallel, so they could

not completely isolate the effect of the model and the scheduling tools). The implementation maintained service levels, while reducing overtime; and through better staff planning and scheduling, staff can usually work 5 days a week rather than the traditional 6 days. Since staff reimbursement is mainly determined by the number of EOBs processed, this does not affect the staff income, but improves their work-life balance. The scheduling tools also help the company to prepare ahead for busy shifts and to better plan staff transportation (the second highest operating expense, after labor).

Potential research directions include incorporating a demand forecast by client group into the model. This would enable the company to schedule employees with the specific skills required to process arriving demand from certain clients at different times, rather than using minimum demand coverage constraints as discussed in Section 3.4.1. This line of research would also require the development of techniques to improve the accuracy of the demand forecast at the client level, such as including information about the EOBs that are already in the 'pipeline' and will arrive soon. Another related research direction is to consider a stochastic version of WDSM that considers different demand scenarios and potential forecast errors.

CHAPTER IV

REUSABLE RESOURCE CAPACITY PLANNING FOR SCHEDULED DEMAND UNDER MINIMUM SERVICE CONSTRAINTS

4.1 Introduction

In this research, we consider a resource planning problem that arises in systems with non-consumable resources, which do not diminish after use and their capacity becomes available at a later time period, possibly after going through a recovery or preparation process (e.g., cleaning). Demand consists of a set of jobs, where each job has a scheduled start time and a duration of the service. There are multiple job types, each corresponding to a particular demand class and requiring a predefined subset of resources. A job is either 'accepted' (i.e, processed/serviced) or 'rejected' (i.e., lost or goes through an alternative channel by outsourcing, rescheduling to a future time, etc.). The service level is defined as the (weighted) proportion of jobs accepted. While more resources (inventory) could lead to higher service levels, additional resources come at a cost (investment, storage, maintenance, etc.). Hence, the goal is to balance the cost of building this capacity and the service level.

This problem is motivated by hospital operations, in particular, by the surgical instruments planning, where most of the surgical cases (jobs) are scheduled in advance according to surgeons' and patients' preferences and staffed operating rooms (ORs) availability. The average U.S. hospital has an inventory of surgical instruments worth approximately between $2 and $4 million dollars, but very few of these institutions have adequate systems for their planing and management, and there are significant opportunities for improvement and cost reduction [64]. Bachmann et al.

[10] report savings of $31 per surgical case by only improving one OR's handling of reusable gynecologic laparoscopy equipment, including promoting instrument accessibility, eliminating infrequently used instruments on permanent trays, etc.

Similar problems also arise in service industries where there is a schedule of projects to complete, workforce of different skills is needed to complete each specific project, and the workforce manager needs to plan for personnel with the right characteristics to assign to each project. Other examples include repair and maintenance applications, where an adequate inventory of repair tools and equipment is needed for scheduled repair and maintenance jobs.

We propose deterministic and stochastic models for finding the optimal capacity level for each resource type, minimizing the cost of resource inventories, subject to service level constraints. These service constraints are considered at both 'global' and demand-class levels, the latter motivated by service contracts with some clients. In Section 4.2 we introduce relevant literature. In Section 4.3 we introduce the problem and the nomenclature and in Section 4.3.1 we present results on the problem's complexity. A deterministic model to solve this problem is presented in Section 4.4. Even though jobs' characteristics such as start and service times are assumed fixed, in some applications, such as the OR surgical instruments planning, resource planning decisions are often done before having full information on the upcoming demand schedule(s). In these cases, resource capacity decisions should be *robust* (i.e, be able to provide certain service metrics considering different demand scenarios). In Section 4.5 we introduce a stochastic model to address the robustness of the resource capacity decisions under different demand schedules. We propose a Sample Average Approximation (SAA) solution approach in Section 4.5.1, and in Section 4.5.2 we show its convergence to the original stochastic model. In Section 4.6 we present the results of a case study based on surgical data of a community hospital to gain insights on the

effect of different model parameters and settings on metrics such as the surgical instruments cost and service levels. Finally, in Section 4.7 we present some conclusions of this research and propose potential research directions.

4.2 Literature Review

The setting described in this research has similarities to the resource-constrained project scheduling problem, which is defined by a set of activities that must be scheduled, subject to precedence and resource constraints, such that the makespan is minimized [72]. In contrast, in our setting the schedule of the activities is fixed, and the focus is on the resource capacity decisions. Also relevant to our setting are resource levelling problems [98] where resource planning decisions are considered with the objective of minimizing capacity investment cost, such that all the activities are completed; unlike our setting which includes accept/reject decisions for jobs and service constraints.

Slotnick [109] presents a review of the research on order (job) acceptance and scheduling, which considers the trade-offs of accepting business and its associated costs of processing. Our setting is closer to the project selection and scheduling problem (PSS) [36, 83, 87]. PSS is reduced to a scheduling problem when all projects must be accepted, whereas if the jobs' schedule is fixed, as it is the case in our setting, PSS is equivalent to a knapsack problem [106]. Weingartner and Ness [123] introduce the multidimensional 0-1 knapsack problem to solve a capital budgeting problem with project selection, given expenditure limitations in several time periods and/or several inputs. Freville [63] presents an overview of the multidimensional 0-1 knapsack problem literature, which, regarding exact solutions, has more commonly been studied as a special case of a 0-1 mathematical program. However, different from our problem, the resource capacities are usually considered as constraints, not decisions, and the objective is often to maximize the value of the chosen projects,

rather than maintaining a service level while minimizing the resource costs.

Both demand acceptance/rejection and resource capacity decisions are relevant in the tool-kit planning problem, where the goal is to find an optimal kit of parts or tools to perform a subset of jobs (e.g., repairs). The choice of a kit involves evaluating the trade-offs between holding costs and the penalty for failing to complete the jobs. Brumelle and Granot [26] study the tool-kit planning problem where each job requires at most one unit of a given resource, and different from our problem, without minimum service level constraints. Teunter [112] studies a version of the problem where repair parts (consumable resources) are required for several job requests, before the kit is restocked. The authors do consider a service-oriented goal (minimum holding cost for a required job-fill rate). Our problem differs from the tool-kit problem in that we consider the timing of the jobs and accept/reject decisions. Hence, we need to not only determine the quantity of each resource, but also their schedule. Ceran et al. [34] study a related problem where software elements are used in different software jobs and analyze trade-offs between additional development time to make these elements 're-usable' and time savings that result from their re-use. The sequence of the jobs is considered, but since each software element can be re-used several times, there are no inventory (resource capacity) decisions.

In settings with orders with stochastic arrival and/or duration times, each order's accept/reject decision is usually taken after each order's arrival (i.e., it is dynamic). Kleywegt and Papastavrou [80] introduce a dynamic knapsack problem with Poisson arrivals. Balakrishnan et al. [13] study the capacity allocation problem faced by make-to-order (MTO) manufacturing under a two-class random demand and one shared resource, resulting in selective rejection of orders for the class with the lower unit contribution. Herbots et al. [73] consider both order selection and resource scheduling decisions, and the number of resources available at each time. The authors derive a stochastic dynamic program for order selection, with equal interarrival times

for all orders and only one type of resource. Similar to the traditional knapsack problem applications, the goal is to find policies for order selection that maximize profits given capacity constraints rather than a balanced multiple resource inventory to achieve a given service level, which is the case in our setting.

Maintaining certain service levels is the goal of many inventory problems under stochastic demand systems. An example of such systems is a loss network, where arriving jobs of different classes, each class simultaneously requiring different set of resources, are lost if they cannot be fulfilled with the resources on-hand. However, computing fill rates (i.e., the percentage of demand covered with the available resources) in the case of multiple demand classes and resource types is hard, even under strong assumptions such as exponential inter-arrival and service times [78, 88]. Order accept/reject decisions are introduced as admission control policies. Nevertheless, for most practical systems it is prohibitively difficult to compute optimal policies since the number of structured policies grows exponentially, so heuristics with guaranteed performance are usually proposed [60, 85, 99, 103]. In other systems, demand is assumed to *always* be fulfilled, either from available stock or from an 'emergency channel', and the service level is given by the percentage of demand fulfilled from stock [118]. Our problem is similar to that described by Güllü and Köksalan [66]. They propose an optimization model and a heuristic procedure to find stock levels that minimize costs of surgical instruments and implants for orthopedic surgeries, subject to a service level. Nevertheless, in this and the other stochastic models, strong assumptions regarding the demand arrival (usually assumed to be Poisson) are needed; whereas in our problem, demand comes from a schedule which should be taken into account.

To the best of our knowledge, the resource planning problem we study for scheduled demand under minimum service constraints has not been previously considered in the literature. Our contributions include the introduction and complexity analysis

of this problem, a mathematical model for its solution, and a convergence analysis of an approximation of an stochastic extension, as well as a case study based on a practical application and real data to better understand the effect of different model parameters and settings.

4.3 Problem Description

We consider a set of jobs. At the beginning of the planning horizon, information about their scheduled start time (arrival), duration (service time), and demand class (type) is available. There is also a set of resources of different types, and a cost for each unit available (inventory) of each type. Based on its demand class, each job requires a set of resources, and it is assigned a weight/reward. We assume that each job can be either accepted or rejected. We define the class service level (CSL) as the fraction of jobs of a given class that are accepted, and we define the global weighted service level (GWSL) as the fraction of jobs that are accepted, weighted according to each job's class.

The problem's notation is as follows:

Sets

J	Jobs
K	Classes of J
I	Resource types
$J_k \subseteq J$	Jobs of class k
$T_j \subseteq J$	Jobs that 'intersect' with job j

Parameters

c_i	Cost per unit of resource type i
t_j	Scheduled start time for job j
d_j	Scheduled duration for job j
$a_{i,j}$	Number of resources of type i required by job j
r_j	Relative weight (value) of job j, where $\sum_{j \in J} r_j = R$ is the total number of *weighted* jobs.

β $\in [0,1]$. Minimum GWSL required
β_k $\in [0,1]$. Minimum CSL required

K only includes those demand classes represented in J; therefore, $|J_k| > 0$ for all $k \in K$. Each job belongs to one (and only one) class, and every job of the same class k requires the same set of resources and has the same weight, i.e., $a_{i,p} = a_{i,q}$ and $r_p = r_q$ if $p, q \in J_k$. We assume that the cost per unit of resource c_i, the resource requirements $a_{i,j}$, and the job relative weights r_j are integer. d_j represents the time between the start time t_j and when the resources used for job j become available again. Then, if $t_q \geq t_p + d_p$, all the resources used for job p can be used (if needed) for job q. If job j is accepted, *all* the resources required by job j must be available from t_j until $t_j + d_j$. T_j represents the jobs that continue or end during the interval $(t_j, t_j + d_j)$. Figure 20 shows an example with six jobs, two demand classes, and three resource types.

The objective of this problem is to find the optimal quantity of each resource type and the job accept/reject decisions that minimize cost, such that the targets for GSWL and CSLs can be achieved.

4.3.1 Special Cases and Complexity

Figure 21 shows 12 different special cases of the problem according to the characteristics of the system. For example $1/I/ \leq 1/1/0$ represents the case with one class, where each job requires at most 1 unit of each resource $i \in I$, all jobs need to be accepted ($\beta = 1$), and there are no restrictions for the CSLs ($\beta_k = 0$, $k \in K$).

Proposition 4.3.1. *Special cases of $K/I/a_{i,j}/\beta/\beta_k$ which are polynomially solvable or NP-complete are shown in Figure 21.*

The proofs can be found in Appendix I.

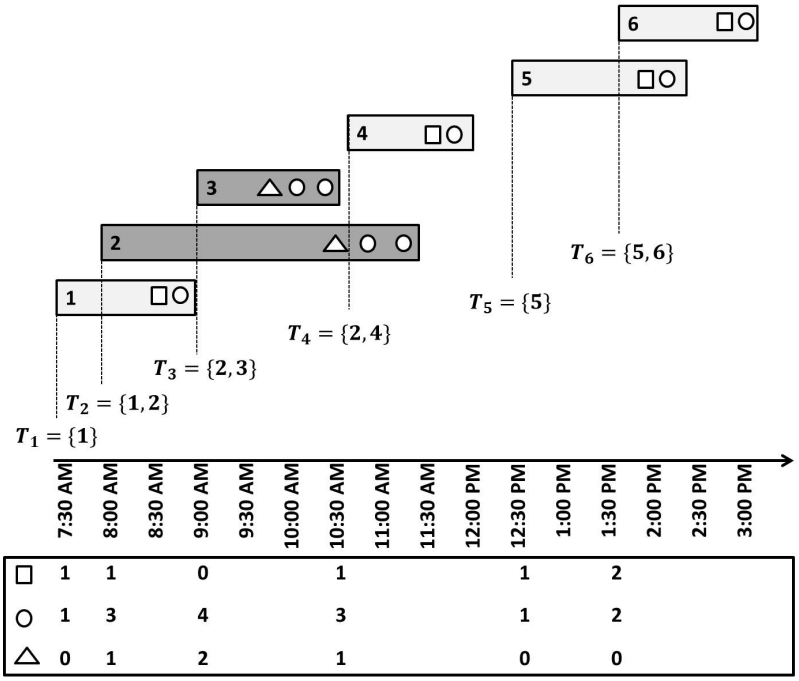


Figure 20: Example of six jobs to complete, with two demand classes, and three resource types, showing intersecting jobs sets T_j for each job and the number of resources required at each job's start time.

$\|K\|$	$\|I\|$	$a_{i,j}$	*All Jobs Accepted*	β *(Global Weighted Service Level Restriction)*	β_k *(Class Service Level Restrictions)*
1	*≥1*	*≥0*	$1/I/a_{i,j}/1/0$	$\mathbf{1/I/a_{i,j}/\beta/0}$	
≥1	*1*	*1*	$K/1/1/1/0$	$\mathbf{K/1/1/\beta/0}$	$\mathbf{K/1/1/0/\beta_k}$
≥1	*1*	*≥0*	$K/1/a_{i,j}/1/0$	$\mathbf{K/1/a_{i,j}/\beta/0}$	$K/1/a_{i,j}/0/\beta_k$
≥1	*≥1*	*≤1*	$K/I/\leq 1/1/0$	$\mathbf{K/I/\leq 1/\beta/0}$	$K/I/\leq 1/0/\beta_k$
≥1	*≥1*	*≥0*	$\mathbf{K/I/a_{i,j}/1/0}$	$K/I/a_{i,j}/\beta/0$	$K/I/a_{i,j}/0/\beta_k$

Polynomially solvable NP-Complete

Figure 21: Different cases of the problem in terms of demand classes, resource types and requirements, and service level restrictions.

4.4 Resource Planning Model (RPM)

We propose a Mixed Integer Program (MIP) to solve the resource planning problem under scheduled demand and service constraints. The decision variables are: (i) quantity of resource $i \in I$, f_i, and (ii) job $j \in J$ accept/reject decision, x_j, where $x_j = 1$ if job j is accepted and 0 otherwise.

The mathematical formulation for this resource planning model (RPM) is as follows:

$$MIN \quad Cost^{RPM} = \sum_{i \in I} c_i f_i \tag{66}$$

$$\sum_{q \in T_j} a_{i,q} x_q \leq f_i \quad i \in I,\ j \in J \tag{67}$$

$$\frac{1}{R} \sum_{j \in J} r_j x_j \geq \beta \tag{68}$$

$$\frac{1}{|J_k|} \sum_{j \in J_k} x_j \geq \beta_k \quad k \in K \tag{69}$$

$$f_i \in \mathbb{Z}_{\geq 0} \quad i \in I \tag{70}$$

$$x_j \in \{0, 1\} \quad j \in J \tag{71}$$

Objective function (66) minimizes the total cost of the available resources. Constraints (67) ensure that the requirements of all accepted jobs do not exceed the available resources (see Figure 20 for an example). There are two types of service constraints: (i) the 'global' service constraint (GSC), represented by constraint (68),

and (ii) 'class' service constraints (CSCs), represented by constraints (69). Constraints (70) and (71) are the non-negative integral and binary constraints for the resource quantities and the accept/reject decisions, respectively.

We can easily construct an explicit schedule for *each* resource from a RPM solution. Let $J^* = \{j \in J : x_j = 1\}$ be the set of jobs to accept, sorted in *ascending order* by their start time t_j, and there is a 'virtual' queue i of the f_i available resources of type i. Assign resources to each job $j \in J^*$, one by one, by first releasing and sending to the corresponding queue all type i resources previously assigned to earlier jobs that have already become available at time t_j, and then assigning the first $a_{i,j}$ resources from queue i to job j. We know there are enough resources in each queue by constraints (67).

Note that it is easy to consider an initial available resource inventory f_i^0 of resource i, by adding constraints (72) to RPM:

$$f_i \geq f_i^0 \quad i \in I \tag{72}$$

4.5 Stochastic Resource Planning Model (SRPM)

RPM assumes that the demand schedule is known at the time when the resource planning decisions are made. While this may be the case in some applications, consider the case where capacity decisions need to be made in advance, and once the demand schedule is revealed, the jobs' accept/reject decisions can be made. In this case, it is desirable to make the resource planning decisions such that they are robust under different demand schedules.

Let W be the (finite) set of possible demand schedules (scenarios), with each schedule $w \in W$ having a probability P^w. We extend the notation introduced in Section 4.3. J^w is the set of jobs in demand schedule w, where t_j^w, d_j^w, $a_{i,j}^w$ and r_j^w are the start time, duration, resource requirements, and weight of job $j \in J^w$. Similarly,

R^w is the number of weighted jobs, and J_k^w is the set of jobs of class k in demand schedule w. In the same fashion, T_j^w is set of jobs in demand schedule w that intersect with job $j \in J^w$. Finally, $x_j^w = 1$ if job j of the demand schedule w is accepted, and 0 otherwise. In addition, let $W_k \subseteq W$ be the set of demand schedules with a least one job of class k (i.e., $|J_k^w| > 0$), and P_k^w be the probability of demand schedule w conditional to $w \in W_k$.

We propose a stochastic resource planning model (SRPM) to minimize the total cost of the available resources, such that the expected global and per-class service levels are at least β and β_k for class k, respectively. The model is as follows:

$$MIN \quad Cost^{SRPM} = \sum_{i \in I} c_i f_i \tag{73}$$

$$\sum_{q \in T_j^w} a_{i,q}^w x_q^w \leq f_i \quad i \in I,\ j \in J^w,\ w \in W \tag{74}$$

$$\sum_{w \in W} P^w \left[\frac{1}{R^w} \sum_{j \in J^w} r_j^w x_j^w \right] \geq \beta \tag{75}$$

$$\sum_{w \in W_k} P_k^w \left[\frac{1}{|J_k^w|} \sum_{j \in J_k^w} x_j^w \right] \geq \beta_k \quad k \in K \tag{76}$$

$$f_i \geq f_i^0 \quad i \in I \tag{77}$$

$$f_i \in \mathbb{Z}_{\geq 0} \quad i \in I \tag{78}$$

$$x_j^w \in \{0, 1\} \quad j \in J^w,\ w \in W \tag{79}$$

4.5.1 Sample Average Approximation (SAA) for solving SRPM

Even when the set of possible schedules W is finite, it may be very large. In that case, solving SRPM optimally might be prohibitively time consuming. In addition, the probability P^w of each schedule $w \in W$ might not be easy to estimate. SAA is an approach commonly used to overcome these difficulties. The main idea is to generate a random sample of W to approximate the original stochastic model and solve this approximate model. Let $S \subseteq W$ be an independently and identically distributed (i.i.d.) random sample of demand schedules, and let $S_k \subseteq S$ be the subset of schedules that contain at least one job of class k ($w \in W : |J_k^w| > 0$). The SAA for SRPM can be written by replacing W by S in SRPM, and modifying GSC (75) and CSCs (76) as follows:

$$\frac{1}{|S|} \sum_{w \in S} \left[\frac{1}{R^w} \sum_{j \in J^w} r_j^w x_j^w \right] \geq \beta \tag{80}$$

$$\frac{1}{|S_k|} \sum_{w \in S_k} \left[\frac{1}{|J_k^w|} \sum_{j \in J_k^w} x_j^w \right] \geq \beta_k \quad k \in K : |S_k| > 0 \tag{81}$$

We call this approximate model Sample Average Approximation Resource Planning Model (SAARPM).

4.5.2 Convergence of SAARPM

Traditional stochastic programming models involve expectation only in the objective function. It has been shown that for given models, the SAA method converges to the solution of the original problem exponentially fast with the sample size [42, 81]. Ahmed and Shapiro [3] extend these results to stochastic programs with integer recourse (i.e, with integer second-stage decisions). Wang and Ahmed [122] find similar convergence results for expected value single-constraint problems. Branda [25] derives an exponential converge rate for problems with mixed-integer solutions and multiple expected-value constraints.

Wang and Ahmed [122] and Branda [25] do not explicitly consider per-scenario constraints and recourse decisions, which are common in two-stage stochastic programs such as SPRM. Anitescu and Birge [8] present a framework for ensuring convergence of stochastic programs with both expectation and per-scenario constraints, using Lagrangian relaxation. However, they assume that both first stage and recourse decision variables are continuous and the model's functions differentiable, and they do not derive a rate of convergence. Next, we extend the results of Wang and Ahmed [122] (summarized in Appendix J) to SAARPM, which has second-stage decision variables and both per-scenario and expectation constraints.

Let F be the set of feasible decisions for the capacity decision variables $\mathbf{f} = (f_i,\ i \in I)$. Note that $\mathbf{f}$ could be bounded below by the initial inventory $\mathbf{f}^0 = (f_i^0,\ i \in I)$, and bounded above by the number of resources required to complete *all* jobs for any demand schedule $w \in W$, when the number of jobs and their resource requirements are bounded above for any schedule.

Let $X(\mathbf{f})$ be the feasible region for the accept/reject decisions x_j^w, for *each* job $j \in J^w$ and demand schedule $w \in W$, given $\mathbf{f}$ available resources and *without* considering service level constraints (i.e., GSC and CSCs),

$$X(\mathbf{f}) = \left\{ \mathbf{x} \in \{0,1\}^{\sum_{w \in W} |J^w|} : \sum_{q \in T_j^w} a_{i,q}^w x_q^w \leq f_i \quad i \in I,\ j \in J^w,\ w \in W \right\} \tag{82}$$

Consider a perturbation vector of size $|K| + 1$, $\Delta = (\epsilon, \epsilon_k\ k \in K)$, $\epsilon > 0$ and $\epsilon_k > 0$, $k \in K$. Let $F^{+\Delta}$ be feasible region of $\mathbf{f}$ for SPRM($+\Delta$), which corresponds to SPRM as described in Section 4.5, except that constraint (75) is substituted by constraint (83), and constraints (76) are substituted by (84),

$$\sum_{w \in W} P^w \left[\frac{1}{R^w} \sum_{j \in J^w} r_j^w x_j^w \right] \geq \beta + \epsilon \tag{83}$$

$$\sum_{w \in W_k} P_k^w \left[\frac{1}{|J_k^w|} \sum_{j \in J_k^w} x_j^w \right] \geq \beta_k + \epsilon_k \quad k \in K \tag{84}$$

Then, F^0 is the feasible region of SPRM. Similarly, we define $F^{-\Delta}$ as the feasible region of $\mathbf{f}$ for model SPRM($-\Delta$), where constraint (75) is substituted by constraint (85), and constraints (76) are substituted by (86),

$$\sum_{w \in W} P^w \left[\frac{1}{R^w} \sum_{j \in J^w} r_j^w x_j^w \right] \geq \beta - \epsilon \tag{85}$$

$$\sum_{w \in W_k} P_k^w \left[\frac{1}{|J_k^w|} \sum_{j \in J_k^w} x_j^w \right] \geq \beta_k - \epsilon_k \quad k \in K \tag{86}$$

SPRM($+\Delta$) and SPRM($-\Delta$) are a *more* and a *less restrictive* versions of SPRM, respectively. Let $F^{|S|}$ be the feasible region of $\mathbf{f}$ for model SAAPRM with a sample S of schedules. The goal is to estimate a lower bound for $P(F^{+\Delta} \subseteq F^{|S|} \subseteq F^{-\Delta})$, i.e., the probability that a feasible solution for SAAPRM with a sample S is Δ-feasible for SPRM, and at the same time it is not too conservative (i.e., having a much larger inventory/cost).

4.5.2.1 Convergence in the Case of One Expectation Constraint

We first consider a special case for SPRM where there are no CSCs (i.e., $\beta_k = 0$ for all $k \in K$), so that the only non-trivial expectation constraint is GSC (75). We define $F^{+\epsilon}$ ($F^{-\epsilon}$) as the feasible region of $\mathbf{f}$ for SPRM($+\delta$) (SPRM($-\delta$)), where $\delta = (\epsilon, 0\, k \in K)$.

Let $L^{max}(\mathbf{f}, w)$ be the maximum GWSL (i.e., the maximum global weighted service level) attainable under demand schedule $w \in W$, given available resources $\mathbf{f}$,

$$L^{max}(\mathbf{f}, w) = \max_{\mathbf{x} \in X(\mathbf{f})} \left(\frac{1}{R^w} \sum_{j \in J^w} r_j^w x_j^w \right) \tag{87}$$

Given $\mathbf{f}$ resources, we can accept or reject jobs so that GWSL is maximized for each demand schedule w. In fact, there is a vector $\mathbf{x} \in X(\mathbf{f})$ where the resulting GWSL for each schedule w is the maximum attainable given $\mathbf{f}$. Then, $l^{max}(\mathbf{f})$ is the expected maximum GWSL, attainable given available resources $\mathbf{f}$,

$$l^{max}(\mathbf{f}) = \sum_{w \in W} P^w L^{max}(\mathbf{f}, w) \tag{88}$$

Both $L^{max}(\mathbf{f}, w)$ and $l^{max}(\mathbf{f})$ can only take values in $[0, 1]$ for all $\mathbf{f} \in F$ and $w \in W$. Then, $l^{max}(\mathbf{f})$ is well-defined, and the MGF of $L^{max}(\mathbf{f}, w) - l^{max}(\mathbf{f})$ is finite (i.e, assumptions (C2) and (C3) in [122] hold; see Appendix J).

Given a sample S of demand schedules $w \in W$, $l^{max_{|S|}}(\mathbf{f})$ is the sample average maximum GWSL, given available resources $\mathbf{f}$,

$$l^{max_{|S|}}(\mathbf{f}) = \frac{1}{|S|} \sum_{w \in S} L^{max}(\mathbf{f}, w) \tag{89}$$

Proposition 4.5.1. *Consider the case with no CSCs. Suppose that that F is a nonempty compact set. Then, given $\epsilon > 0$, $P(F^{+\epsilon} \subseteq F^{|S|} \subseteq F^{-\epsilon})$ converges to 1 exponentially fast as the sample size $|S|$ increases and,*

$$P(F^{+\epsilon} \subseteq F^{|S|} \subseteq F^{-\epsilon}) \geq 1 - 2|F|e^{-\frac{|S|\epsilon^2}{2\sigma^2_{Lmax}}}$$

Where σ^2_{Lmax} is the maximum variance of $L^{max}(\boldsymbol{f}, w) - l^{max}(\boldsymbol{f})$ among all $\boldsymbol{f} \in F$, i.e., $\sigma^2_{Lmax} = \max_{\boldsymbol{f} \in F} Var[L^{max}(\boldsymbol{f}, w) - l^{max}(\boldsymbol{f})]$.

Moreover, since $P(F^{+\epsilon} \subseteq F^{|S|} \subseteq F^{-\epsilon}) \geq 1 - 2|F|e^{-\frac{|S|\epsilon^2}{2\sigma^2_{Lmax}}}$, it follows that a lower bound for the sample size $|S|$, such that $P(F^{+\epsilon} \subseteq F^{|S|} \subseteq F^{-\epsilon}) \geq 1 - \alpha$, is given by the inequality $|S| \geq \frac{2\sigma^2_{Lmax}}{\epsilon^2} \ln(\frac{2|F|}{\alpha})$.

The proof is shown in Appendix K.

4.5.2.2 Convergence in the Case of Multiple Expectation Constraints

We now consider the case where in addition to GSC (75) we have one or more CSCs (76). Without loss of generality, let K be the set of demand classes with a corresponding non-trivial CSC, i.e., $k \in K : \beta_k > 0$.

Let $L(\mathbf{f}, \mathbf{x}, w) \in [0, 1]$ be the GWSL obtained under demand schedule $w \in W$, given accept/reject decisions $\mathbf{x} \in X(\mathbf{f})$:

$$L(\mathbf{f}, \mathbf{x}, w) = \frac{1}{R^w} \sum_{j \in J^w} r_j^w x_j^w \tag{90}$$

Then, $l(\mathbf{f}, \mathbf{x}) \in [0, 1]$ is the expected GWSL given available resources $\mathbf{f}$ and accept/reject decisions $\mathbf{x} \in X(\mathbf{f})$:

$$l(\mathbf{f}, \mathbf{x}) = \sum_{w \in W} P^w L(\mathbf{f}, \mathbf{x}, w) \tag{91}$$

Similarly, let $L_k(\mathbf{f}, \mathbf{x}, w) \in [0, 1]$ be the CSL (i.e, class service level) for demand class k obtained under demand schedule $w \in W_k$ (where $W_k \subseteq W$ is the subset of demand schedules with at least one job of class k), given accept/reject decisions $\mathbf{x} \in X(\mathbf{f})$:

$$L_k(\mathbf{f}, \mathbf{x}, w) = \frac{1}{|J_k^w|} \sum_{j \in J_k^w} x_j^w \tag{92}$$

$l_k(\mathbf{f}, \mathbf{x}) \in [0, 1]$ is the expected CSL for class k given available resources $\mathbf{f}$ and accept/reject decisions $\mathbf{x} \in X(\mathbf{f})$:

$$l_k(\mathbf{f}, \mathbf{x}) = \sum_{w \in W_k} P_k^w L_k(\mathbf{f}, \mathbf{x}, w) \tag{93}$$

Both $l(\mathbf{f}, \mathbf{x})$ and $l_k(\mathbf{f}, \mathbf{x})$ are well-defined, and the MGFs of $L(\mathbf{f}, \mathbf{x}, w) - l(\mathbf{f}, \mathbf{x})$ and $L_k(\mathbf{f}, \mathbf{x}, w) - l_k(\mathbf{f}, \mathbf{x})$ are finite.

Given a sample S demand schedules $w \in W$, we define $l^{|S|}(\mathbf{f}, \mathbf{x}) \in [0, 1]$ as the sample average GWSL, given accept/reject decisions $\mathbf{x} \in X(\mathbf{f})$:

$$l^{|S|}(\mathbf{f},\mathbf{x}) = \frac{1}{|S|}\sum_{w\in S} L(\mathbf{f},\mathbf{x},w) \tag{94}$$

Recall that $S_k \subseteq S$ is the set of sample demand schedules that contain at least one job of class k. Further, sample S is 'class-representative' if all the constrained demand classes $k \in K$ are represented in the sample, i.e., $|S_k| > 0$, $k \in K$ (there is such a sample almost surely with a sample size $|S|$ large enough). Then, $l_k^{|S|}(\mathbf{f},\mathbf{x}) \in [0,1]$ is the sample average CSL for $k \in K$, $|S_k| > 0$, given second-stage decisions $\mathbf{x} \in X(\mathbf{f})$:

$$l_k^{|S|}(\mathbf{f},\mathbf{x}) = \frac{1}{|S_k|}\sum_{w\in S_k} L_k(\mathbf{f},\mathbf{x},w) \tag{95}$$

Proposition 4.5.2. *Suppose that that F is a nonempty compact set, and that sample S is class-representative. Then, given $\Delta = (\epsilon, \epsilon_k\ k \in K)$, $P(F^{+\Delta} \subseteq F^{|S|} \subseteq F^{-\Delta})$ converges to 1 exponentially fast as the sample size $|S|$ increases and,*

$$P(F^{+\Delta} \subseteq F^{|S|} \subseteq F^{-\Delta}) \geq 1 - 2(|K|+1)|F|e^{-|S|\min\{\frac{\epsilon^2}{2\sigma^2}, \frac{\epsilon_k^2}{2\sigma_k^2}\ k\in K\}}$$

where

$$\sigma^2 = \max_{\boldsymbol{x}\in X(\boldsymbol{f}),\boldsymbol{f}\in F} Var[L(\boldsymbol{f},\boldsymbol{x},w) - l(\boldsymbol{f},\boldsymbol{x})],\ \ \sigma_k^2 = \max_{\boldsymbol{x}\in X(\boldsymbol{f}),\boldsymbol{f}\in F} Var[L_k(\boldsymbol{f},\boldsymbol{x},w) - l_k(\boldsymbol{f},\boldsymbol{x})]$$

Moreover, since $P(F^{+\Delta} \subseteq F^{|S|} \subseteq F^{-\Delta}) \geq 1-2(|K|+1)|F|e^{-|S|\min\{\frac{\epsilon^2}{2\sigma^2}, \frac{\epsilon_k^2}{2\sigma_k^2}\ k\in K\}}$, it follows that a lower bound for the sample size $|S|$ such that $P(F^{+\Delta} \subseteq F^{|S|} \subseteq F^{-\Delta}) \geq 1-\alpha$ is given by the inequality $|S| \geq \frac{1}{\rho}\ln(\frac{2(|K|+1)|F|}{\alpha})$, where $\rho = \min\{\frac{\epsilon^2}{2\sigma^2}, \frac{\epsilon_k^2}{2\sigma_k^2}\ k \in K\}$.

The proof is shown in Appendix L. The bounds in Proposition 4.5.1 and Proposition 4.5.2 show (at least) an exponential convergence of the feasible region of SAARPM. The lower bound in Proposition 4.5.2 can be also used in the case where there is only one expectation constraint (e.g., GSC). However, Proposition 4.5.1 offers a better bound since $\sigma_{Lmax}^2 \leq \sigma^2$. While σ^2 represents the largest variance between a service level (weighted or per class if there is one CSC in lieu of GSC) for a random

demand schedule and the expected service level, across all possible capacity decisions (i.e, for all $\mathbf{f} \in F$) and all resulting feasible accept/reject decisions (i.e., for all $\mathbf{x} \in X(\mathbf{f})$); σ^2_{Lmax} is the largest variance across all possible capacity decisions and a vector of accept/reject decisions (say $\mathbf{x}^{Lmax}$) that *maximizes* the service level for each scenario $w \in W$, which is also feasible (i.e, $\mathbf{x}^{Lmax} \in X(\mathbf{f})$).

4.6 Case Study

This research was motivated by the planning of surgical instruments. With the exception of emergency and add-ons, surgical cases are scheduled in advance based on the surgeons' and patients' preferences and the staffed ORs availability. Each surgical case requires instruments and specialized instrument trays depending on the procedure to be performed, and in some cases, on the surgeon performing the case. A preference card is a list including all the materials (both consumable and non-consumable) that are required for the procedure. Before the surgical case is scheduled to start, the listed materials are picked-up and put in a cart -the case cart- and set-up in the OR during the *room turnover*, which includes all the room cleaning and preparation activities. After the surgical case ends, the case cart is brought to the cleaning and sterilization area, where the cart and its contents are washed and decontaminated. Then, the instruments are regrouped into trays, wrapped, and put through the sterilization process. After the sterilization process is completed, the instruments are left to cool off. Once the instruments cool off, they are either put in the next case cart to be set-up in an OR for another surgical case or stored in the storage area (clean core). The cycle time of the instruments decontamination, wash, sterilization, and cooling adds to the surgical case set-up and are included in the total service time in our computational experiments.

During the surgical cases scheduling, it is assumed that the required instruments

are always available (with exceptions of larger equipment such as portable body scanners). If some instruments required for a surgical procedure are not available when they are needed (i.e., at the start of the surgery time), an emergency case cart may be used or the surgery may have to be rescheduled. Flash sterilization can also be used to speed the sterilization process and increase the availability of instruments, but this is not recommended as adverse events have been related with this practice [33]; failure to properly sterilize surgical instruments carries a high risk related to the breach of the patient's physical and immune defense, leading to an infection [104]. Therefore, improving surgical instruments planning and management can not only bring economical but also health-related benefits.

In this setting, there are important decisions regarding the planning of resources (surgical instruments), such that cost is minimized and the required service levels are met. The proposed deterministic model, RPM, addresses these resource planning decisions under the assumption that the surgical schedule is known in advance and some adjustments can be done when the capacity decisions are made. Closer to reality is SAAPRM, an approximation of SPRM, where resource planning decisions are made well in advance, and different demand schedules are then revealed and adjustments can be made regarding which jobs (surgeries) will be performed as planned with the available instruments (i.e., 'accepted' jobs) versus which may require the utilization of emergency carts or rescheduling (i.e., 'rejected' jobs).

We implement RPM and SAAPRM using historical surgical data from a 580-bed, not-for-profit, community hospital, which performs about 8,000 surgical procedures annually in 14 ORs. Next, we describe how we use these surgical data to generate surgical schedules as an input to our models, as well as the assumptions made on the surgical instruments requirements.

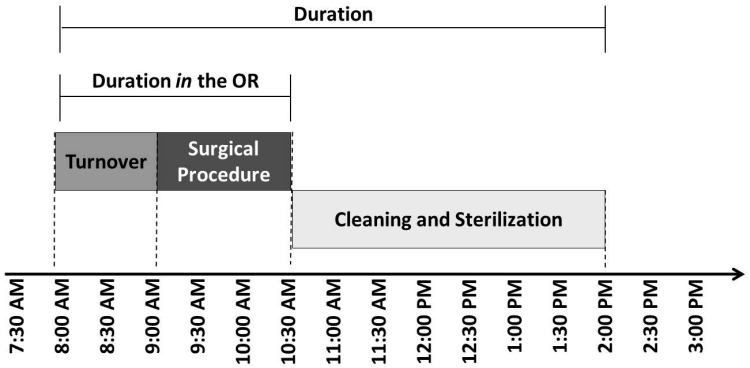


Figure 22: Example of the duration of a surgical instrument cycle.

4.6.1 Data Analysis and Generation of Surgical Schedules

To implement RPM and SAARPM, we need information regarding each job's arrival, duration or service time, and demand class or type. In this application, each job corresponds to a surgical case, composed of one (e.g., total knee replacement) or more (e.g., cystoscopy and insertion of ureteral stent) surgical subprocedures. Each surgical procedure, i.e., the combination of one or more subprocedures, corresponds to a demand class and requires specific surgical instruments. These instruments must be available in the OR from the start of the room turnover to the end of surgery, after which the instrument cleaning and sterilization process begins. An example of a surgical instrument cycle is shown in Figure 22. In this example, the surgical instruments should be available in the OR at 8:00AM to initiate the room turnover for the next surgical procedure, which ends at 10:30AM. The instruments become available again at 2:00PM after completing the cleaning and sterilization process (including cooling). Hence, the total duration of this 'job' (corresponding to the surgical case) is of 6 hours, considering cleaning and sterilization.

We assume that there is a minimum inventory (initial inventory f_i^0) to cover any surgical case alone. In addition, we assume that the OR manager can 'reject' (i.e, use emergency carts, or cancel or reschedule to a later time) some cases after the surgical schedule is revealed to avoid schedule conflicts regarding the instruments' availability.

We assume that *the surgical schedule is known for (at least) two weeks in advance.*

We analyze about four years of surgical data (January, 2009 to November, 2012). There are around 1,300 different subprocedures, and more than 4,000 different procedures scheduled during this time, but 387 of these 4,000 account for 80% of the cases. Based on each case's *surgical service* (group of surgeons that perform surgical procedures of a specific specialty), we classify the procedures in 13 specialties: cardiothoracic, colon-rectal, general, gynecology, neurosurgical, oral and maxillofacial, orthopedic, otorhinolaryngology, plastic surgery, urology, vascular, and other. We assign each of the most common 387 procedures to a different demand class, and we group the rest in 13 classes corresponding to each of the 13 specialties, for a total of 400 procedure classes.

This surgery department uses both open access and *block schedule*. The block schedule assigns a time block of a specific OR on a given day of the week to a particular surgical service(s). For this reason, similar cases are more likely to be scheduled in the same OR around the same time and day of the week. To account for this, we classify each of the more than 29,000 cases in the surgical data by OR, day of the week, and sequence order in the OR (e.g., OR 1, Monday, second case of the day). We also compute the proportion of 'closed' days (i.e., when no cases are scheduled) for each OR on each day of the week (e.g., 10% for OR 2 on Tuesdays).

The surgical data includes the scheduled start and end times for each case. We assume that the room turnover should start when the previous case ends, which is consistent with the current practice. However, we also assume a limit of 90 minutes for room turnover between cases as this is long enough for an OR cleaning and set up, even for those ORs with slow turnover times [47]. Therefore, if in the surgical data a case is scheduled to start more than 90 minutes after the previous case scheduled end time, we assume that there is an intentional 'schedule gap', and the case cannot be scheduled earlier (for instance, if the surgeon is not available), and we only allocate

90 minutes before the case for turnover. For an OR's first case of the day, we assume that the room turnover takes 30 minutes (as the current practice). Following these assumptions, we compute the scheduled turnover start time and duration in OR (room turnover plus surgical procedure) for each case in the surgical data.

In addition to the scheduled start and end times, the surgical data includes the actual start (wheels-in) and end (wheels-out) times for each case. When we compare the actual and the scheduled case duration in the OR, we observe that often the scheduled time underestimates the actual time (10 minutes, on average, for the fours years of available surgical data). We also found deviations when we compare the actual and the scheduled room turnover start times (including only those cases when the OR's previous case ends on-time, indicating another source of delay). Using the surgical data we find empirical distributions for deviations in the scheduled case duration in the OR, the scheduled turnover start time for a first case and for a second or later case, for each of the 13 procedures specialties. Examples of this analysis are presented in Appendix M.

After the case ends, all the surgical instruments go through a cleaning and sterilization process. The sterilization time depends on the technology used, which can vary by the type of instrument; for example, urology instruments take about 15 hours to clean and sterilize whereas orthopedic instruments take about 3 hours and 45 minutes. For this reason, the cleaning and sterilization time is considered by procedure specialty. Most of cleaning and sterilization process is automatic, so there is not much duration variability (therefore, considered as constant). For this hospital, the cleaning and sterilization process takes (on average) about 60% of the instrument service cycle.

We generate surgical schedule scenarios based on the available surgical data, one day, one OR at a time, in the following manner:

1. Randomly determine if the OR is open for the day (given the proportion of days

closed for each OR/week day combination). If the OR is closed, we move to the next OR.

2. Select a random first case among those scheduled in the same day of the week, OR, and sequence order in the surgical data.

   (a) Based on the data, define the scheduled start and end times (including room turnover).

   (b) Add a random 'noise' to the start time and duration in the OR, based on the deviation empirical distributions for the selected procedure specialty.

   (c) Add the cleaning and sterilization time based on the procedure specialty.

3. If the selected first case in step 2 is followed by a second case in the data, randomly select a case among those scheduled in the same day of the week, OR, and sequence order in the data. Otherwise, we move to the next OR.

   (a) Schedule the case right after the first case (scheduled) end, unless there is a schedule gap. In this case, the surgical case is scheduled based on the data. The duration in the OR is specified by the data.

   (b) Follow sub-steps (b) and (c) from step 2.

4. If the selected second case is followed by a third case, repeat the step 3. Otherwise, move to the next OR.

We repeat the process above for each OR, for each day in the planning horizon. The random selection of the cases from the surgical data can be uniform (i.e, all cases scheduled on the same week day, OR and order have the same chance) or more recent cases can have a higher chance to be selected. Figure 23 shows an example of the scheduling of an OR's second case 'Case B'. Case B is scheduled in the given OR-day right after Case C at 9:30AM (note that it is the second case of the day). However,

Table 18: Comparison of randomly generated weekly surgical schedules and surgical data.

	Surgical Data	Randomly Generated	t-test p-value
Average Number of Cases/Week	136.67	136.71	0.97
Average Case Duration	7:24	7:26	0.31
Average First OR Case Start Time	7:59	7:58	0.40
Average Last OR Case End Time	20:11	20:19	<0.00

given the adjustments that come from the deviation empirical distributions, Case B actually starts 30 minutes after Case C (which also was delayed) ends at 11:00AM, and lasts 30 minutes longer than scheduled. Figure 24 shows an example where there is a schedule gap between surgical data's Case A and Case B. For this reason, Case B is not scheduled to start until 11:30AM (as it is in the surgical data). We still assume there is a start time delay of 30 minutes.

We compare relevant metrics of the weekly schedules in the surgical data (199 weeks, after eliminating the fist and last week of each year from the sample) and those generated by the process described above (with a sample of 530 weeks), assuming that the random selection of the cases is done uniformly. The average and t-test p-values are shown in Table 18. There is not a statistically significant difference in the average number of cases per week, the average duration, and the average first OR case start time. There is a statistically significant difference in the average last OR case end time, but it is only of 8 minutes. This suggests that we are able to approximate the hospital's surgical scheduling process. In the computational experiments (Section 4.6.2), we used a triangular distribution $P(z) = (z-1)^2/(n-1)^2$ for the random selection of cases (rather than a uniform distribution), where n is the number of cases to choose from, and $\lfloor z \rceil$ represents the (rounded to nearest integer) order of the selected case in the list, ordered by increasing start time and date.

We consider three groups of surgical instruments (or instrument trays): (i) general instruments used in different procedure specialties (e.g., some cutting instruments),

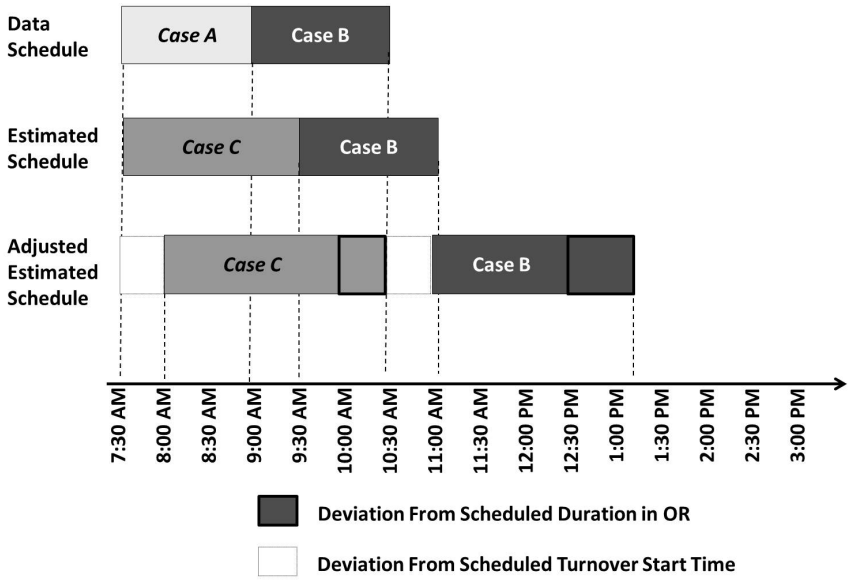


Figure 23: Example of scheduling an OR's second case.

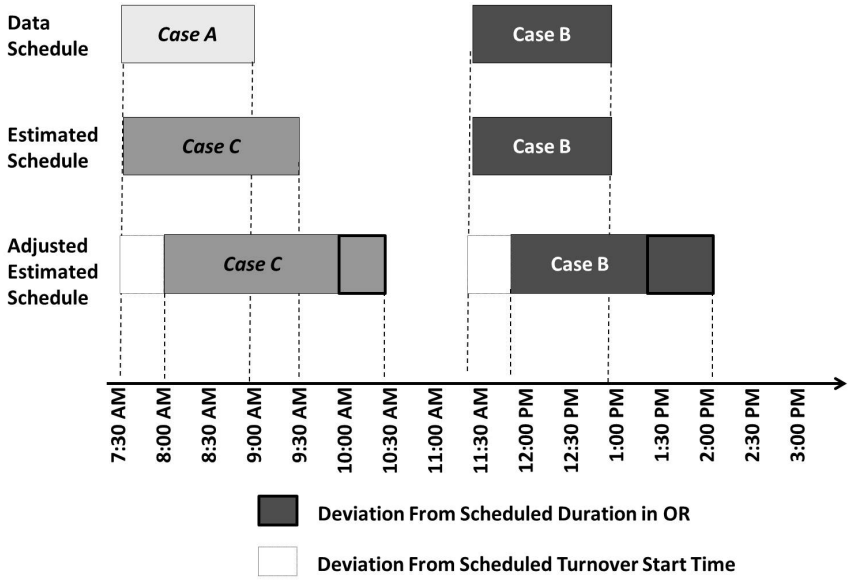


Figure 24: Example of scheduling an OR's second case with a gap in the schedule.

Table 19: Probability of requiring 0, 1, 2, or 3 units of each instrument.

Instrument	P(0)	P(1)	P(2)	P(3)
General A	0.1	0.4	0.4	0.1
General B	0.3	0.4	0.3	0
General C	0.6	0.3	0.1	0
General D	0.5	0.5	0	0
General E	0.75	0	0.25	0
General F	0.8	0.15	0.05	0
General G	0.75	0.25	0	0
General H	0.875	0	0.125	0
General I	0.9	0.07	0.03	0
General J	0.95	0.03	0.02	0
Specialty 1-13	0.1	0.4	0.4	0.1
Specialty 14-26	0.3	0.4	0.3	0
Specialty 27-39	0.6	0.3	0.1	0
Specialty 40-52	0.8	0.15	0.05	0
Specialty 53-65	0.9	0.07	0.03	0
Specialty 66-78	0.95	0.03	0.02	0

(ii) specialty instruments used only in procedures of a specific specialty (e.g, orthopedic pliers), (iii) subprocedure instruments used in specific subprocedures (e.g., a mammotome used in breast biopsies). We assume that each procedure class requires 0, 1, 2 or 3 units of the 10 general instruments and 0, 1, 2, or 3 units of the 6 instruments by procedure specialty according to a given probability (see Table 19). Also, each procedure class requires 1 unit of the subprocedure-specific instrument corresponding to each one of its subprocedures. There are 10 general instruments, 78 specialty instruments, and 1,334 subprocedure instruments, for a total of 1,422 instrument types (corresponding to the set of resource types I).

4.6.2 Computational Experiments

We run a series of experiments to gain insights on the effects of different model parameters and settings on metrics such as cost and service levels. We also compare the deterministic and stochastic resource planning models. We are interested in answering the following questions:

1. For RPM, what is the effect of:

   (a) Having a shorter or longer planning horizon on the surgical instruments'

cost?

(b) The instrument group (general, specialty or subprocedure-specific) on the instrument inventory growth versus the minimum service level required?

(c) Surgical procedure characteristics such as duration, complexity, and frequency on the service level per class (i.e., surgical procedure)?

(d) Introducing CSCs in RPM on the cost and the service level for surgical procedures with unconstrained service levels (i.e, class k such that $\beta_k = 0$)?

(e) Increasing the 'weights' assigned to some surgical procedures. How does this compare to introducing CSCs for these procedure classes?

(f) Reducing the surgical instrument cycle duration e.g., by introducing a new technology that would allow to *safely* reduce the cleaning and sterilization time?

(g) The accept/reject decisions given a fixed inventory for each resource (instrument)?

2. When using SAAPRM to model the capacity decisions made in advance:

(a) How does the capacity cost convergence under different sample sizes of surgical schedule scenarios?

(b) What is the value of knowing the surgical schedule at the time of making the capacity planning decisions?

To answer these questions, in our computational study we use the experimental settings described in Table 20. We classify the surgical procedures (demand classes) based on their frequency: 'high', scheduled five or more times every two weeks ($K_h \subseteq K$, $|K_h| = 7$); 'medium', scheduled at least once every two weeks but less than five times ($K_m \subseteq K$, $|K_m| = 43$); and 'low', scheduled less than once every two weeks ($K_l \subseteq K$, $|K_l| = 350$). The goal is to study the effect of using CSCs or larger weights

Table 20: Experimental settings.

	Model	Sample Size ($\|S\|$)	Horizon (weeks)	$\beta \cdot 100\%$	$\beta_k \cdot 100\%$	Weight (r_j)	Sterilization
I	RPM	1	{2, 4, 12, 24, 48}	{10%, 30%, 50%, 70%, 90%, 95%, 99%}	None	Same	Current
II	RPM	1	{2, 4, 12, 24, 48}	{10%, 30%, 50%, 70%, 90%, 95%, 99%}	90% $k \in K_h$, 85% $k \in K_m$, 0% $k \in K_l$	Same	Current
III	RPM	1	{2, 4, 12, 24, 48}	{10%, 30%, 50%, 70%, 90%, 95%, 99%}	None	3 $k \in K_h$, 2 $k \in K_m$, 1 $k \in K_l$	Current
IV	RPM	1	{2, 4, 12, 24, 48}	{10%, 30%, 50%, 70%, 90%, 95%, 99%}	None	Same	75% longer
V	RPM	1	{2, 4, 12, 24, 48}	{10%, 30%, 50%, 70%, 90%, 95%, 99%}	None	Same	25% longer
VI	RPM	1	{2, 4, 12, 24, 48}	{10%, 30%, 50%, 70%, 90%, 95%, 99%}	None	Same	25% shorter
VII	RPM	1	{2, 4, 12, 24, 48}	{10%, 30%, 50%, 70%, 90%, 95%, 99%}	None	Same	75% shorter
VIII	SAARPM	{1, 2, 4, 8, 16, 32, 64}	2	{10%, 30%, 50%, 70%, 90%, 95%, 99%}	None	Same	Current
IX	SAARPM	{1, 2, 4, 8, 16, 32, 64}	2	{10%, 30%, 50%, 70%, 90%, 95%, 99%}	90% $k \in K_h$, 85% $k \in K_m$, 0% $k \in K_l$	Same	Current

for the most common procedures, such as total hip replacement and cardiopulmonary bypass, which represent an important source of revenue and more than 50% of the cases. Moreover, there is a higher risk that common procedures 'intersect' with each other in the schedule, increasing the risk of conflicts regarding the availability of the instruments. There are 9 'settings groups' and a total of 343 different settings (considering the different surgical schedule sample sizes $|S|$, horizon lengths, and the required minimum GWSL, i.e., β). We solve 10 different instances for each setting (for a total of 3,430 instances) using Gurobi 5.6.3. We highlight the main results in the remainder of this section.

4.6.2.1 Effects of the Surgical Schedule Horizon and Surgical Instrument and Procedure Characteristics

To answer questions 1.a, 1.b, and 1.c, we consider experimental setting I (Table 20), where there are no CSCs (i.e, $B_k = 0$ for all $k \in K$) and every procedure has the same weight ($r_j = 1$ for all $j \in J$). We observe that the length of the planning horizon has a significant effect only at higher service levels; as the planning horizon increases, the cost increases in the same direction (see Figure 25). As β increases, most of the cases need to be accepted, and in a longer horizon there are more potential schedule 'conflicts' given the same instrument inventory. On the contrary, with a shorter planing horizon, the inventory mix can be 'specialized'. When analyzing the average inventory per instrument type with respect of β (see Figure 26), we observe that the average inventory grows faster for the general instruments, which are highly shared among all the cases, followed by the specialty instruments. In the case of the

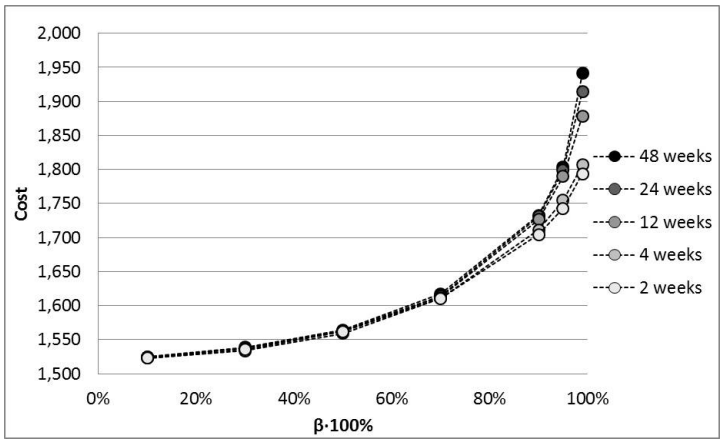


Figure 25: RPM average instruments' cost vs. β and the surgical schedule time horizon.

subprocedure-specific instruments, while few units are required for the most common subprocedures, one unit is enough for most even at high β levels, since it is less likely to have a schedule conflict.

We are also interested in studying the effects of other procedure' characteristics such as duration and complexity in the resulting service level per procedure class (i.e., CSL). Procedures with shorter than average procedure duration are classified as 'short' and 'long' otherwise, and procedures with more than one subprocedure are classified as 'complex' and 'non-complex' otherwise. As expected, cases with shorter and less complex procedures have higher CSLs as they require fewer instruments (see Figure 27); however, the effect is reduced at higher β levels, since there is less flexibility to choose which cases to accept.

4.6.2.2 Minimum Service Levels vs. Different Weights by Surgical Procedure

RPM differentiates the relative importance of surgical procedure (class) k in two ways: (i) by introducing a CSC $\beta_k > 0$ (i.e., a required minimum CSL), and (ii) by assigning a weight r_j to all the surgical cases (jobs) with a procedure k (i.e, $j \in J_k$). We compare settings I, II, and III in Table 20 to answer questions 1.d and 1.e. Table 21 shows the average cost and average CSLs for each of the three procedure groups by

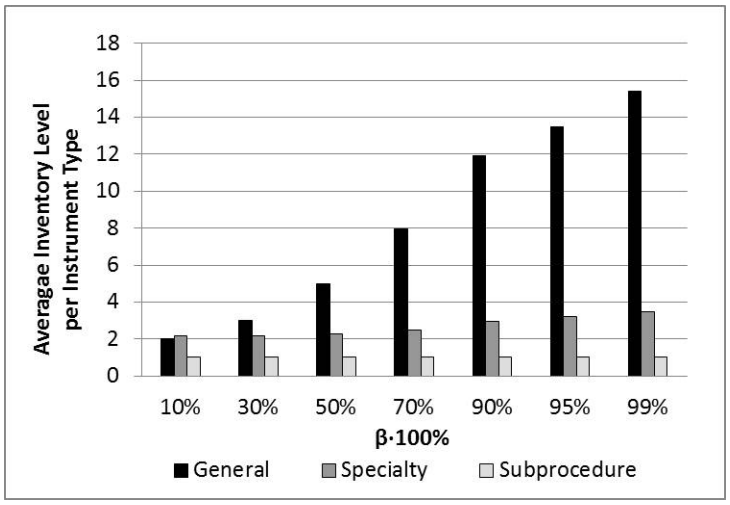


Figure 26: RPM average inventory per instrument type for each surgical instrument group vs. β, given two-week schedules.

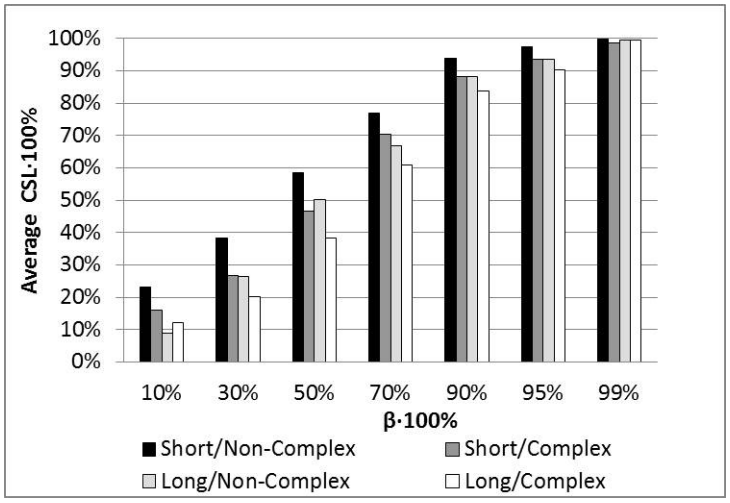


Figure 27: RPM average CSL by procedure for each duration/complexity procedure group vs. β, given two-week schedules.

frequency under a GSC with a β of 90% and a planning horizon of two and 48 weeks. The planning horizon affects not only the average cost, but also the resulting average CSLs. In setting I, with two-week schedules, higher frequency procedures have lower service levels. This makes sense since more frequent cases have a higher probability of conflicting instrument schedules, so they are rejected in a larger proportion. However, with 48-week schedules, the higher frequency procedures have a higher CSL, since increasing the horizon allows for the instruments required by these procedures to be reused more often, and thus allowing certain specialization.

When we introduce CSCs with $\beta_k > 0$ for the high and medium frequency surgical procedures, the average CSL increase for these procedures. The increase is more dramatic in the case of the two-week schedules, because the 'round-up' of the minimum number of cases to accept (i.e, $\lceil N_k \cdot \beta_k \rceil$) has more impact, e.g., if there are 5 cases and we need to accept 90% of them, we need to accept at least $\lceil 4.5 \rceil = 5$. The average CSLs for the low frequency procedures is not considerably affected because decreasing too much the service level for these procedures would affect the GWSL and we have that $\beta = 90\%$. When we compare introducing CSCs for high and medium frequency procedures to increasing their cases' weights, we observe that whereas these procedures classes result with high average CSLs, the low frequency procedures are penalized. Increasing the weight for certain procedures allows for specialization, so that higher CSLs can be achieved for these procedures, without increasing the cost (or even reducing it) while achieving the same GWSL. Using higher weights for some procedures and introducing CSCs for others can be done simultaneously to specialize or give the inventory a focus, while guaranteing minimum service levels for particular procedures.

Table 21: RPM average cost and CSL by procedure frequency group under (procedure-based) CSCs and different procedure weights, given settings I, II, and III, two-week and 48-week schedules, and $\beta = 0.90$.

	Setting I		Setting II		Setting III	
	Two-week	48-week	Two-week	48-week	Two-week	48-week
$\beta_k \cdot 100\%$ Weight (r_j)	None Same		K_h: 90%, K_m: 85% Same		None K_h: 3, K_m: 2, K_l: 1	
Average Cost	1,704	1,732	1,717	1,736	1,690	1,708
Average CSL, High Frequency ($k \in K_h$)	89%	93%	98%	93%	97%	97%
Average CSL, Medium Frequency ($k \in K_m$)	90%	87%	100%	90%	95%	90%
Average CSL, Low Frequency ($k \in K_l$)	92%	89%	82%	87%	76%	77%

Table 22: Change of the RPM average inventory per instrument type under different sterilization methods, by instrument group, given two-week schedules and $\beta = 0.90$.

Sterilization Method	General	Specialty	Subprocedure
75% Longer	16.7%	5.2%	0.2%
25% Longer	4.4%	1.9%	0.1%
25% Shorter	-11.5%	-3.4%	-0.2%
75% Shorter	-24.5%	-10.6%	-0.5%

4.6.2.3 Effect of Different Sterilization Technologies

As discussed in Section 4.6.1, the technology used in the cleaning and sterilization process has an important impact in an instrument's cycle duration. We analyze the effect of alterative technologies with shorter and longer times, by reducing and increasing the reported cleaning and sterilization time (question 1.f). We compare experimental settings I, IV, V, VI, and VII (Table 20). The average inventory of general and specialty instruments (most shared) are more affected by a change of technology (see Table 22). The overall reduction in inventory cost at $\beta = 0.90$ is about 4% considering a decease of 75% in the cleaning and sterilization time, but the decrease on the *additional* inventory is about 33% (see Figure 28). Finally, the effect of different cleaning and sterilization times is only significant at higher β levels.

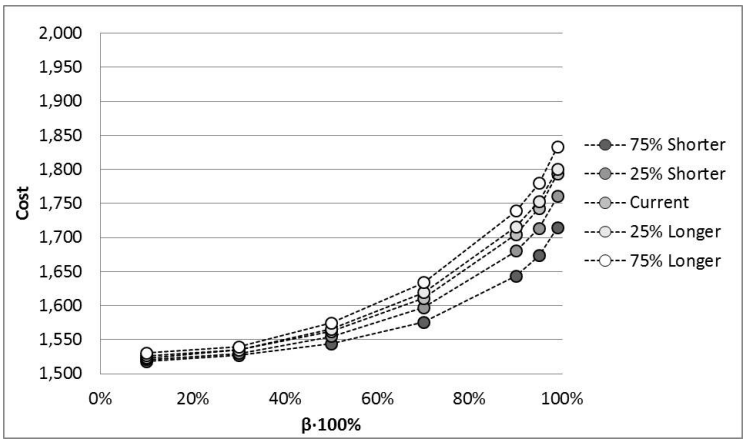


Figure 28: RPM average instruments' cost vs. β and different sterilization methods.

4.6.2.4 Impact of the Accept/Reject Decisions Under Fixed Resource Capacities

RPM finds the optimal cases (jobs) accept/reject decisions to minimize the cost of the required instruments to complete them. To answer question 1.g, we study the effect of implementing different (and potentially not optimal) aception/rejection decisions on service levels, given a fixed instrument inventory. We assume a first-in selection heuristic (FISH): we accept any case (ordered by start time) unless we cannot fulfill all its requirements with the available instruments at the case's start time. This is equivalent to the OR manager 'doing nothing' regarding the accept/reject decisions. To test this heuristic, we solve RPM under Table 22 settings I and II (RPM with and without CSCs, respectively) and obtain the optimal quantity for each instrument i. We implement FISH using these capacity results and compute the resulting GWSL and CSLs. We still assume that we know the surgical schedule by the time the capacity decisions are made, so when implementing FISH we use the same surgical schedule scenario used when solving RPM.

Figure 29 shows the average GWSL obtained using FISH, compared with the minimum required, i.e., β. Even when the surgical case acception/rejection decisions are not optimal, the capacity decisions provide service levels close to β. This is particularly true with high β values and longer planning horizons. Note that it is possible

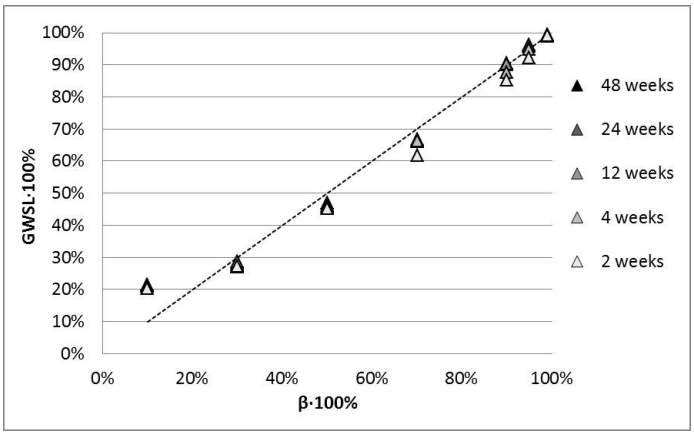


Figure 29: Average FISH GWSL and β vs. β, given two-week schedules, given RPM and setting I capacity decisions.

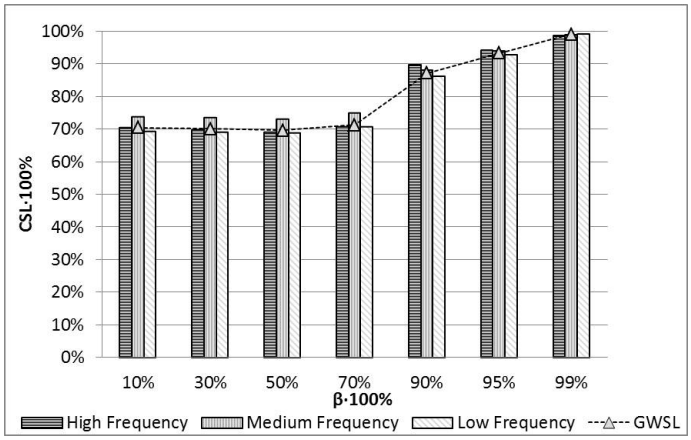


Figure 30: Average FISH CSLs by procedure frequency groups vs. β, given RPM and setting II capacity decisions.

for GWSL to be larger (but no smaller) than β in RPM. However, in Figure 30, the average CSL fall behind the minimum required in setting II for high and medium frequency procedures (90% for $k \in K_h$ and 85% for $k \in K_h$, respectively, given setting II). Nevertheless, FISH still works well with higher β values (with respect to β_k values). For instance, at $\beta = 0.90$, the average CSL for high frequency procedures is 90% and 88% for medium frequency. While FISH does not emphasize particular procedures, the capacity decisions resulting from RPM capture in some degree procedure-focused needs through the CSCs.

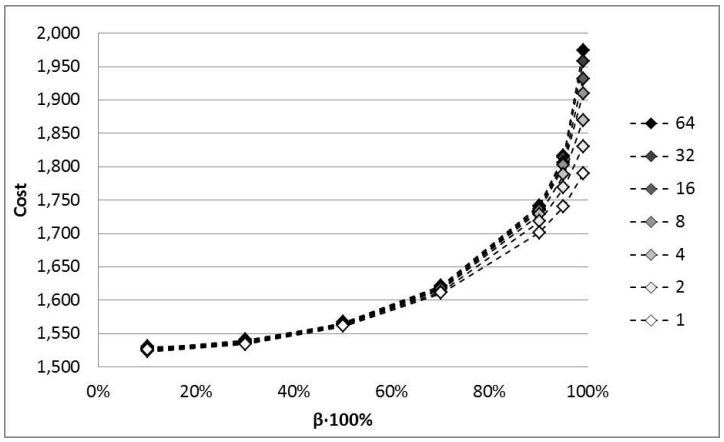


Figure 31: SAARPM average cost vs. β under different sample sizes $|S|$, given setting VIII.

4.6.2.5 SAARPM Convergence

We study SAARPM cost convergence (question 2.a) under Table 22 settings VIII and IX. In the case with no CSCs, the sample size does not have a significant effect at low values of β, which suggests that the convergence for these instance settings is fast (see Figure 31). On the other hand, the solution shows more variability at higher β values, which slows convergence. For instance, going from $\beta = 90\%$ to $\beta = 99\%$ increases the average cost variability from 3 to 7 (with a sample size $|S| = 68$). At high β values there is also less flexility for accept/reject decisions, so considering a larger sample of schedules can lead to a higher cost (a similar effect than a longer planning horizon in RPM). The effect of a larger sample goes in the opposite direction when SAARPM includes CSCs (setting IX) and under low β values, when GSC is no longer constraining. The average cost is higher for small sample sizes because it is harder to average the low CSL that result from rejecting one or a few cases of a given procedure in a smaller sample size (see Figure 32).

4.6.2.6 The Value of the Surgical Schedule Information

SAARPM finds one inventory solution that minimizes instrument inventory cost but at the same time works well (on average) for a sample of surgical schedules, while RPM

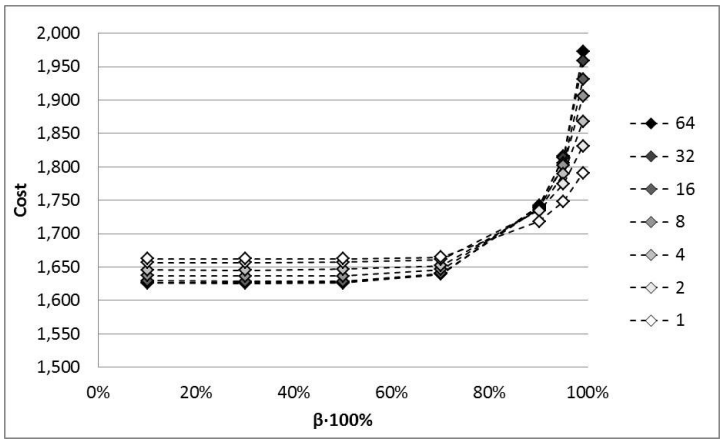


Figure 32: SAARPM with CSCs average cost vs. β under different sample sizes $|S|$, given setting IX.

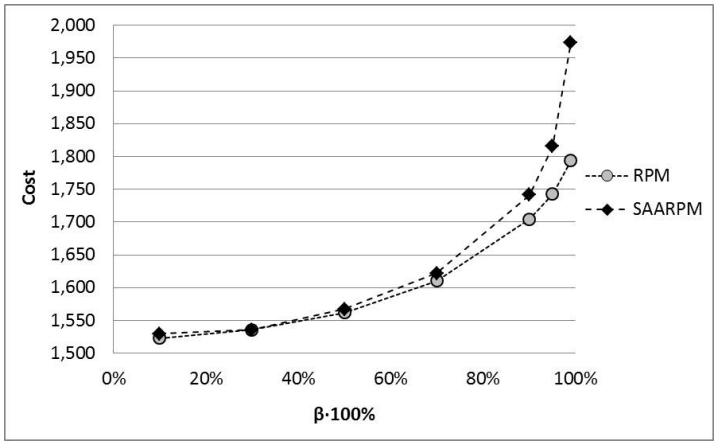


Figure 33: SAAPRM ($|S| = 68$) and RRM (two-week schedule) average costs vs. β.

finds one inventory solution that works well for one particular schedule. To answer question 2.b, we compare the average cost that results from settings I and VIII in Table 22. The cost gap is a measure of the value of knowing the surgical schedule when the capacity decisions are made, so that the inventory can be specialized for each particular schedule scenario. Both models' average cost are similar at low values of β; however, there is an increasing gap as β increases, indicating that inventory robustness comes at a higher 'price' as the flexibility of rejecting surgical cases decreases (see Figure 33).

However, SAARPM assumes that optimal accept/reject decisions are implemented

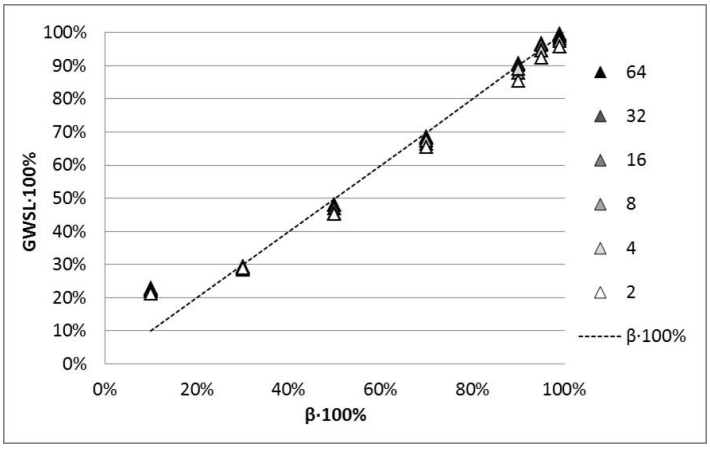


Figure 34: Average FISH GWSL and minimum GWSL β vs. β and sample size $|S|$, given two-week schedules, given SAARPM and setting VIII capacity decisions.

after the surgical schedule information is known. We study the effect on the service levels of implementing FISH for these accept/reject decisions, using the resource capacities solution from SAARPM under settings VIII (no CSCs) and IX (with CSCs). However, different from Section 4.6.2.4, in this analysis we implement FISH using random surgical schedule scenarios, (potentially) different to those used to solve SAAPRM to make the capacity decisions. We find that 'doing nothing' accept/reject decisions can still provide good service levels. In Figure 34, the heuristic's GWSL gets closer to the minimum level required β, as the SAAPRM sample size of surgical schedule scenarios increases (see Figure 34). Similar to the deterministic version, FISH can be adequate in the case of CSCs with $\beta_k > 0$ for some $k \in K$, if β is relatively large, since there is not enough emphasis in the service-constrained procedures otherwise (see Figure 35).

4.7 Conclusions

In this chapter, we introduce a problem of resource capacity optimization under scheduled demand and service constraints. We study the complexity of various special cases of the problem, and propose a optimization model, RPM, to solve it. The model makes two types of decisions: capacity and job accept/reject decisions. RPM

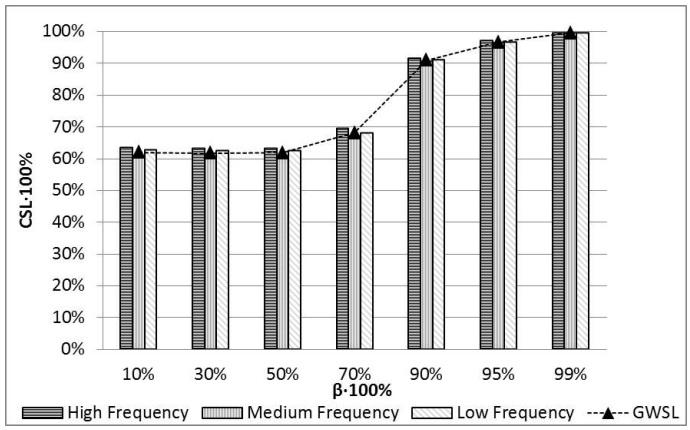


Figure 35: Average FISH CSLs by procedure frequency groups vs. β, given SAARPM and setting IX (with sample size $|S| = 68$) capacity decisions.

assumes that the jobs' schedule is known at the time the capacity decisions are made. While this might the case in some applications, others may require to make these decisions ahead of time to allow enough time for budgeting and procurement activities. For this reason, we propose an stochastic extension of RPM, SRPM. In this stochastic model, capacity decisions are done in order that given service levels can be achieved, on average, across the potential job schedule scenarios. We propose a SAA approach to solve SRPM. We show that this approximation, SAARPM, converges (at least) exponentially fast to the original problem with the scenario sample size.

We also present a case study of the implementation of the proposed models with an application on surgical instruments planning, with the goal of gaining insights about the effects of the models' settings and parameters. Surgical cases (jobs) are scheduled ahead of time in most non-emergency situations using open and/or block scheduling. Each case requires a set of instruments (resources) according to the surgical procedure (demand class), and after each use, the instruments go through a cleaning and sterilization process. We used about four years of surgical data to generate surgical schedule scenarios. Among the main findings, a longer schedule horizon has a significant effect only in high service levels, increasing the cost. The characteristics of the surgical procedure, such as duration and number of subprocedures, have a larger

effect on the procedure's service level in low service levels. We can emphasize surgical cases of a given procedure (i) introducing a CSC and guaranteeing a service level by procedure (at a higher cost), or (ii) assigning a larger weight to these cases, allowing for inventory specialization, with less impact on the cost, but at an expense of other procedures' service levels. We also study the potential effects of new cleaning and sterilization technologies that could allow reduce the instruments' service time. These technology advancements have the most impact in reducing the inventory of resources that are shared across many procedures, and therefore have a higher rotation.

One important assumption of the models is that the surgical schedule must be known by the time of the accept/reject decisions. However, in real world applications the information about the schedule might not be perfect; for instance, there might be unexpected delays, emergencies, etc. This would complicate the implementation of optimal accept/reject decisions. But we found that, even when the models' accept/reject decisions are no implemented as planned, or if there are alternative demand scenarios, the capacity decisions found by RPM or SAARPM can provide service levels close to the minimum required, under a simple surgical case selection process where every surgical case is accepted if all the instruments that are required are available at the start time.

In addition to consider randomness during the accept/reject decisions, problem extensions also include to consider job re-scheduling (e.g., delaying their starting time) rather than cancelling. This would assume that there is certain flexility in the jobs schedule (i.e., a time window for the job start or end times). Another extension might consider that the jobs' durations are not independent, for instant if the resources' recovery process is done in batches. In this case, we would need to also consider the schedule of this recovery process. Nevertheless, even if the problem described does not exactly represent the real world application, the models proposed could still provide a benchmark to evaluate current resource capacities.

APPENDIX A

CHAPTER 2: PHASE II FORMULATION WHEN A STAFF MEMBER CAN BE ASSIGNED TO DIFFERENT SHIFT LENGTHS

In the Phase II original formulation (Section 2.3.2), we assume that a full-time staff member should be assigned to only one type of shift length l, each week. If we only assume that full-time staff should work an average number of hours per week, say H^r, we can drop the index l in $Z_{s,l}$ and introduce the following parameters and variables:

H^r	Average number of regular time hours per week per staff member
Z_s	Number of full-time staff assigned to service line s

We replace constraints (17), (18), and (19) with constraints (96), (97), (98).

$$Z_s \geq \frac{1}{H^r} \sum_{j \in J, d \in D} H_j^l Y_{s,j,d}^{ft} \quad s \in S \tag{96}$$

$$Z_s \geq \sum_{j \in J} Y_{s,j,d}^{ft} \quad s \in S,\ d \in D \tag{97}$$

$$\sum_{j \in J, d \in D} H_j^l Y_{s,j,d}^{pt} + H^r Z_s \leq H^{std} B_s \quad \forall s \in S \tag{98}$$

Constraints (96) are necessary conditions for full-time staff not be scheduled more than H^r hours per week on average, during regular time. This means no more than $\frac{H^r}{8}$ shifts per week, per full-time staff member, since full-time shifts are at least 8 hours long.

APPENDIX B

CHAPTER 2: OTHER MODEL PARAMETERS FOR THE TWO-PHASE ORS STAFFING MODEL

We define the remaining of the parameters for the two-phase ORs staffing model based on the hospital's current processes and practices. These parameters include:

Phase I

- The standard number of hours per FTE per week is $H^{std} = 40$, and the effective number of hours is $H^e = 37.5$.
- Staff hours in excess of 40 hours per FTE per week are paid at an overtime rate. Overtime is paid at a 1.5 rate with respect of regular time, i.e., $C_s^o = 1.5C_s^r$.
- The costs to hire and to fire are $C_s^h = 0$ and $C_s^f = 0$. The initial workforce size is assumed to be zero ($X_s^0 = 0$).

Since we do not have a value for the hospital's maximum fractions of overtime and pooling, we define these parameters so that the resulting permanent FTEs (X_s) would be similar to those implemented by the hospital for the same planning horizon, with the goal of making results more comparable. The baseline for the number of permanent FTEs for both circulators and scrub techs is obtained from the hospital's staffing in December 2012. The staffing budget and structure do not change much throughout a year, so we believe that this is a reasonable approximation of the original staff planning decisions for 2012. Even if this was not the case, Phase I makes decisions by considering about the same or less information (i.e., historical data available prior to July 2011) compared to that available to the hospital's OR manager when the

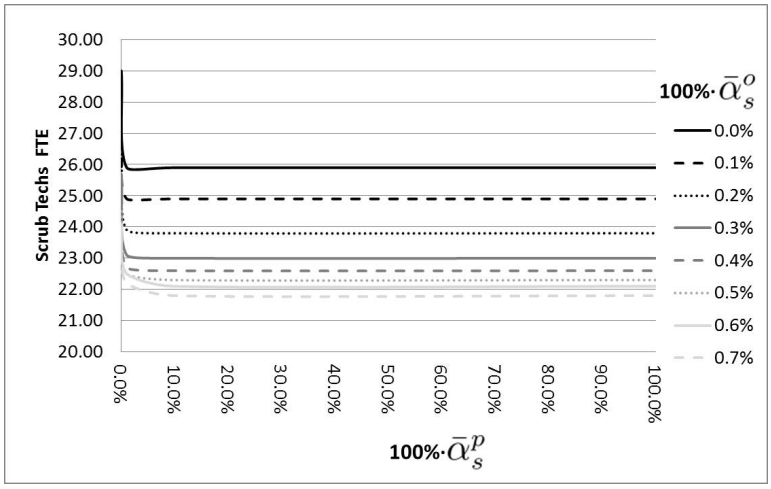


Figure 36: Number of FTEs obtained from Phase I, under different settings for the maximum average percentage of pooling and overtime.

initial staffing decisions are made. In December 2012, the number of FTEs for scrub techs and circulators were 23.8 and 20.7, respectively. We test different maximum pooling and overtime settings to match these numbers of FTEs. Using CPLEX, we run Phase I for circulators with 30 demand scenarios for a 48-week planning horizon in 2012 assuming that no pooling or overtime is allowed. We arrive to approximately the same number of permanent FTEs as the hospital, so we keep the zero pooling and no overtime assumption. Similarly, we run Phase I for scrub techs and look for pooling and overtime settings that would give similar results to those implemented by the hospital. In Figure 36 we see that 23.8 FTEs (the hospital's budget) result from a maximum average overtime ($\bar{\alpha}_s^o \cdot 100\%$) of 0.7% and a maximum average pooling ($\bar{\alpha}_s^p \cdot 100\%$) of 0.0%, or 0.2% and 10% respectively. We select the latter setting, when overtime is lower.

Phase II

- Shifts can start every 30 minutes from 6:00 AM to 2:30 PM, every day of the week. There are five shift lengths: $\{5, 8, 9, 10, 12\}$.

- There is a pre-fixed night shift from 7:00 PM to 6:30 AM with a minimum

staffing requirement of one circulator and two scrub techs. Including this night shift, there are 91 potential shifts ($N_J = 91$).

- The less-than-full-time hours fraction (α_s^{lf}) is limited to 0.04 of effective permanent FTEs hours for scrub techs and zero for circulators.

- The part-time hours fraction (α_s^{pt}) is limited to 0.22 of effective permanent FTEs hours for circulators and scrub techs.

- We use the staffing levels from week 28 in 2010 to week 27 in 2011 as an estimate for the staffing levels during the first 48 weeks of 2012 (the planning horizon).

- The penalty for unmet staffing level by service line is constant for all time buckets and all service lines ($\Pi_{s,t} = 1$).

- An additional 0.5 penalty ($\Pi'_t = 0.5$) is incurred for overall unmet staffing levels.

APPENDIX C

CHAPTER 2: OR SIMULATION

We build the simulation model using Arena. Figure 37 shows a snapshot of the Arena model, with the identification of the main parts:

- Read staff schedule: This module reads a file with the number of available staff at any time for each service line, based on the staffing structure under consideration.
- Read cases information: This module reads a file with the surgical cases information and times.
- Patient arrival and paperwork: This module generates patient arrivals according to the arrival time distributions. After a patient's arrival, paperwork is completed at reception. The patient is prepared for surgery when the estimated procedure start time is closer (< 2 hours).
- Blood drawn and assessment: Blood may be drawn, and the patient is assessed by a nurse.
- Waiting for scheduled time and OR: The patient waits for the scheduled procedure start time if it is the first case for the OR, the anesthesiologist, or the surgeon; otherwise, the patient waits for the OR to be ready.
- Seize anesthesia and OR staff: The case patient seizes the required anesthesia and OR staff. OR staff assigned to the case's service line is chosen first if available, if not, OR staff form other service lines can be pooled and used.

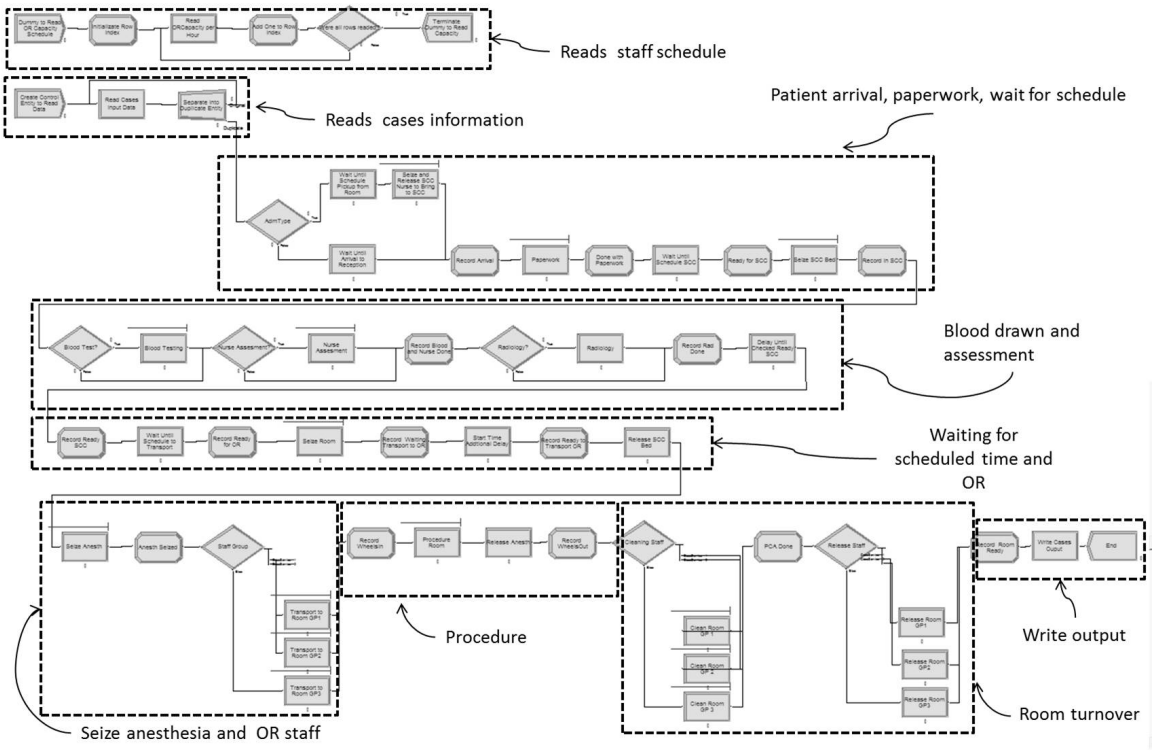


Figure 37: Arena simulation snapshot.

- Procedure: The patient is taken to the OR and the surgical procedure starts. The patient is taken out of the room after the procedure is completed.
- Room turnover: The OR is cleaned and prepared for the next case by the patient care assistants and the OR staff.
- Write output: This module writes the surgical case time stamps in a file for analysis.

To obtain the time distribution for the patient's arrival relative to the case scheduled start time, the percentage of cases with blood being drawn, and the patient assessment and preparation time distributions, we use a 16-days time study done in November 2010 that includes the time stamps of the more than 280 OR patients from their arrival to the end of their preparation for the surgical procedure. We link these time stamps with data provided by the surgical information system, including the

Table 23: Distribution results for OR turnover time.

Case type	Distribution
Cardio	1 - Beta
Open Heart	1 - Erlang(E)
Colon-Rectal	1 - Log-Logistic
General	3 - Log-Logistic(E)
Neuro	1 - Lognormal(E)
Ortho	1 - Pearson Type VI(E)
Other	2 - Log-Logistic(E)
Urology	1 - Johnson SB
Vascular	1 - Beta

type of patient and procedure, the scheduled start and end times for the case, and the patient's wheels-in and -out times. The cases assigned OR, sequence, and actual duration are taken directly from the surgical data. All the time distribution fittings are done using Expert Fit. If there is not a distribution with a good fit, an empirical distribution based on the available data is used. We use empirical time distributions for outpatients and inpatients paperwork, blood work and nurse assessment, and for outpatients arrival time; and we fit a Normal distribution (AD p-value > 0.25) for inpatients arrival time. We use the surgical data from 2009 and 2010 to calculate the turnover time distributions by case type. We assume that the turnover for a case starts at the wheels-out of the previous patient in the room and ends with the wheels-in of the new patient. Only the turnover for cases with some delay are considered, so that the case waiting time for its scheduled start time is not included in the calculation. Table 23 shows the best fitted distributions for each case type. AD p-values > 0.25 and KS p-values > 0.15, for all case types (except for Colon-Rectal, with a KS p-value > 0.10).

C.1 OR Simulation Validation

To validate the simulation, we analyze the simulation results given the hospital's CP staffing structure and compare them with the actual surgical data for the planning horizon, i.e, the first 48 weeks of 2012. In particular, we look at the percentage of delayed cases and an OR last wheels-out time of the day, since these statistics reflect

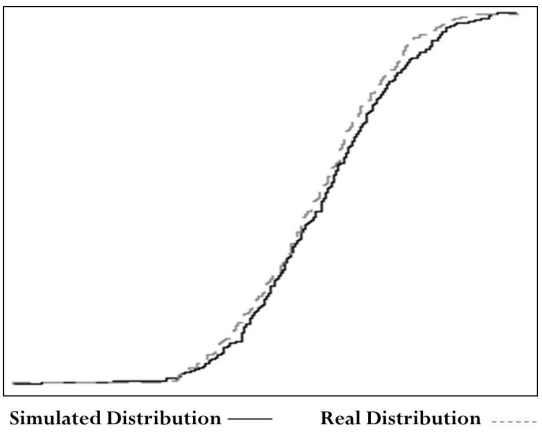


Figure 38: Simulated and actual average OR last wheels-out empirical distribution functions.

the surgical schedule characteristics and the resource availability dynamics.

According to surgical data, the percentage of delayed cases (those cases for which the wheels-in occurs 10 minutes or later after the scheduled time) is 42.4%, whereas according to the simulation this percentage is 41.6%, a non-statistically significant difference of 0.8% (Fisher's exact test p-value > 0.31). The average OR last wheels-out time in a day (for those days with more than 5 ORs open) is 3:37PM according to surgical data and 3:30PM according to the simulation, a non-statistically significant difference of 7 minutes (two-sample t-test with p-vale > 0.65). The average OR last wheels-out time distributions for the actual surgical data and the simulation results are shown in Figure 38. We perform a KS 2-sample test under the null hypothesis that these distributions are not different. According to this test there is not sufficient evidence to reject this hypothesis at a 95% confidence.

APPENDIX D

CHAPTER 2: A DECISION-SUPPORT TOOL

We developed a decision-support tool with the objective of helping the OR manager to make adjustments to the staff budget and the staffing structure by: reassigning staff to another service line, changing the number of people assigned to a shift, adding a new shift, etc. The tool was implemented using Microsoft Excel, and it was automated for ease of use.

Tool Input

- Historical surgical data (from automatic reports from the surgical information system).
- Current staffing structure.
- Classification of surgical services to one of the three service lines.
- OR staff requirements per case type.
- User can select: time baseline, staff type, one or a group of ORs.

Tool Output

- Budget by service line and staff type.
- Weekly volume trends by service line, compared with the budget. See Figure 39.
- Current staffing levels vs. observed demand patterns, overall and by service line. See Figure 40.

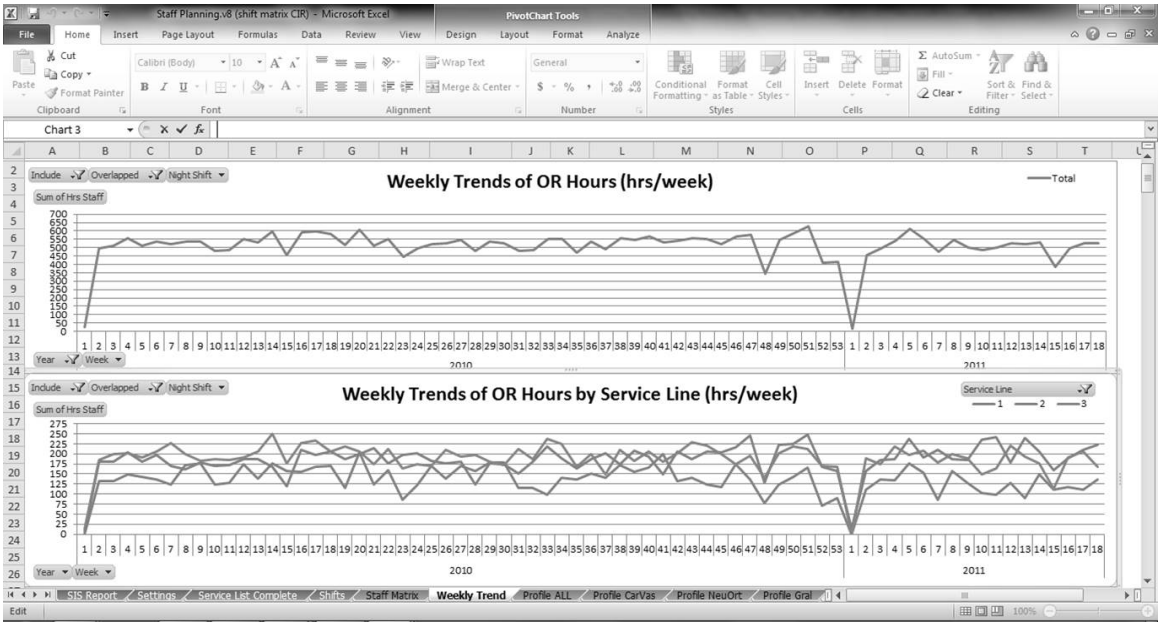


Figure 39: Weekly volume trends by service line in OR staff hours.

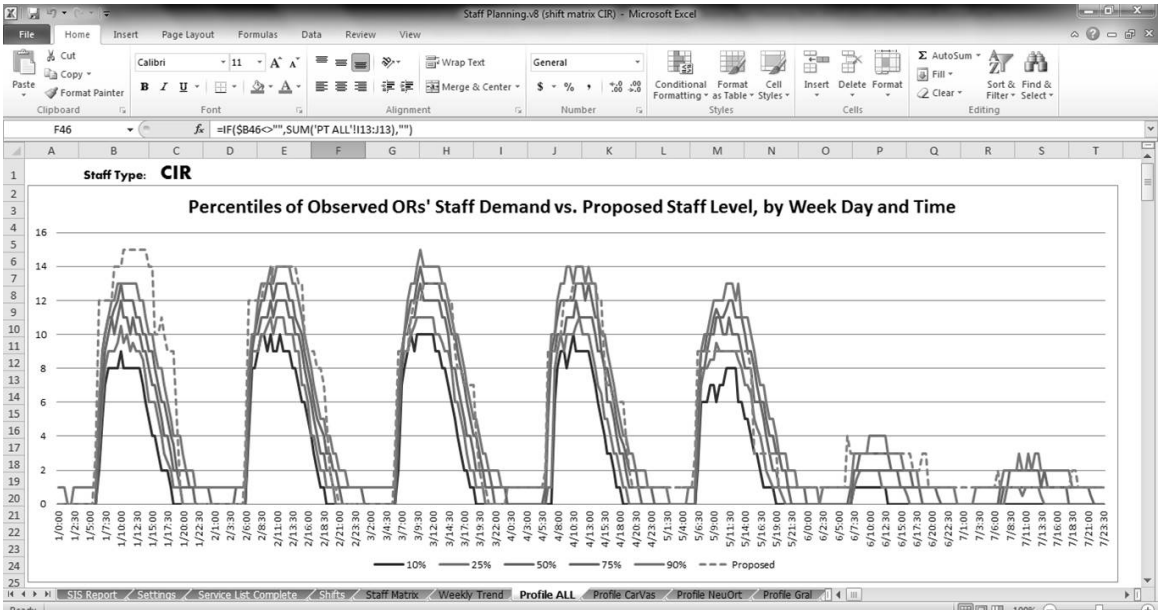


Figure 40: Staffing levels of circulators compared with the demand patterns for all service lines.

Although this tool does not make automated staff planning decisions (it is not integrated with Phase I and Phase II yet), it gives useful statistics and graphs based on easily available surgical data, which are helpful to analyze and fine-tune previous decisions. For example, if the OR manager observes that the demand of a particular service line exhibits a decreasing trend, and the demand of another service line increases, he/she can reassign staff. Similarly, if the aggregated demand trend is increasing, the manager can request an increase in the total FTEs budget. Also, if the demand patterns suggest that a service line may be overstaffed on one day and understaffed on another, the OR manager can restructure shifts.

APPENDIX E

CHAPTER 3: FSM AND THE MINIMUM COST CIRCULATION PROBLEM

Proof. The minimum cost circulation problem is a generalization of the network flow problem. Given a directed graph $G = (V, E)$ (V is the set of nodes and E is the set of arcs), source node $s \in V$ and sink node $k \in V$, where arcs $(u, v) \in E$ have non-negative capacity and a cost per unit of flow; the minimum cost circulation problem consists of finding a flow from s to k that minimizes the total cost. The minimum cost circulation problem is polynomially solvable.

We can model FSM as a minimum cost circulation problem on a directed graph as follows (nomenclature follows Section 3.3):

u_t	'Arriving demand' nodes, $t \in T$
v_t	'Fulfilling demand' nodes, $t \in T : t \geq 1$
s	Source node
k	Sink node
$s'(s'')$	'Virtual' source node for the fraction α of demand arriving in $t \in T : t > 7bN_W - \delta$ that cannot (can) be delayed to the next planning horizon without a penalty
u'	'Virtual' node representing a delay of demand fulfillment to the next planning horizon

Arcs	**Cost**	**Capacity Bounds [lower bound, upper bound]**	**Parameter Ranges**
(u_{t_1}, v_{t_2})	0	$[0, \infty]$	$(t_1, t_2) \in \Delta$
$(v_t, k)_i$	$-\Pi_i^{idl}$	$[0, E_i^{prod}]$	$i \in I, t \in T : X_{i,j,d} = 1, (j, d) \in V_t, t \geq 1$

(s, u_t)	$-\Pi'$	$[0, F_t]$	$t \in T : t \leq 7bN_W - \delta$
(s, s')	$-\Pi'$	$[0, \alpha \sum_{t \in T : t > 7bN_W - \delta} F_t]$	
(s', u_t)	0	$[0, \alpha F_t]$	$t \in T : t > 7bN_W - \delta$
(s, s'')	0	$[0, (1-\alpha) \sum_{t \in T : t > 7bN_W - \delta} F_t]$	
(s'', u_t)	0	$[0, (1-\alpha) F_t]$	$t \in T : t > 7bN_W - \delta$
(s'', u')	0	$[0, (1-\alpha) \sum_{t \in T : t > 7bN_W - \delta} F_t]$	
(u', k)	0	$[0, \infty]$	
(k, s)	0	$[0, \infty]$	

It is easy to see that there is a one-to-one correspondence between the minimum cost circulation problem of this graph and FSM. If there is a positive flow on an arc (u_{t_1}, v_{t_2}), then that corresponds to the value for $A_{(t_1,t_2)}$ in FSM (demand forecasted to arrive at time t_1 and fulfillment scheduled at time t_2). If there is a positive flow on an arc $(v_t, k)_i$, then that flow corresponds to the value for $Y_{i,t}$ in FSM, i.e., this flow equals the demand scheduled to be fulfilled at time t by employee i, who can process up to E_i^{prod} units of demand and is scheduled during time t, since according to the arc definition we have that $X_{i,j,d} = 1$ and $(j, d) \in V_t$. Finally, the forecasted demand F_t minus the flow entering node u_t for $t \leq 7bN_W - \delta$ corresponds to the value for F'_t; and similarly the sum of αF_t for $t > 7bN_W - \delta$ minus the flow on arc (s, s'), corresponds to the value for F''. To compute the cost of FSM, we compute all the penalties given the fixed schedule assuming no demand is fulfilled, and then we add the (non-positive) cost that results from solving the above minimum cost circulation problem.

□

Figure 41 shows the network representation of FSM. In this network, we are assuming that employees i_1 and i_2 are scheduled during time t, and employee i_1 is also scheduled during time $t + \delta$.

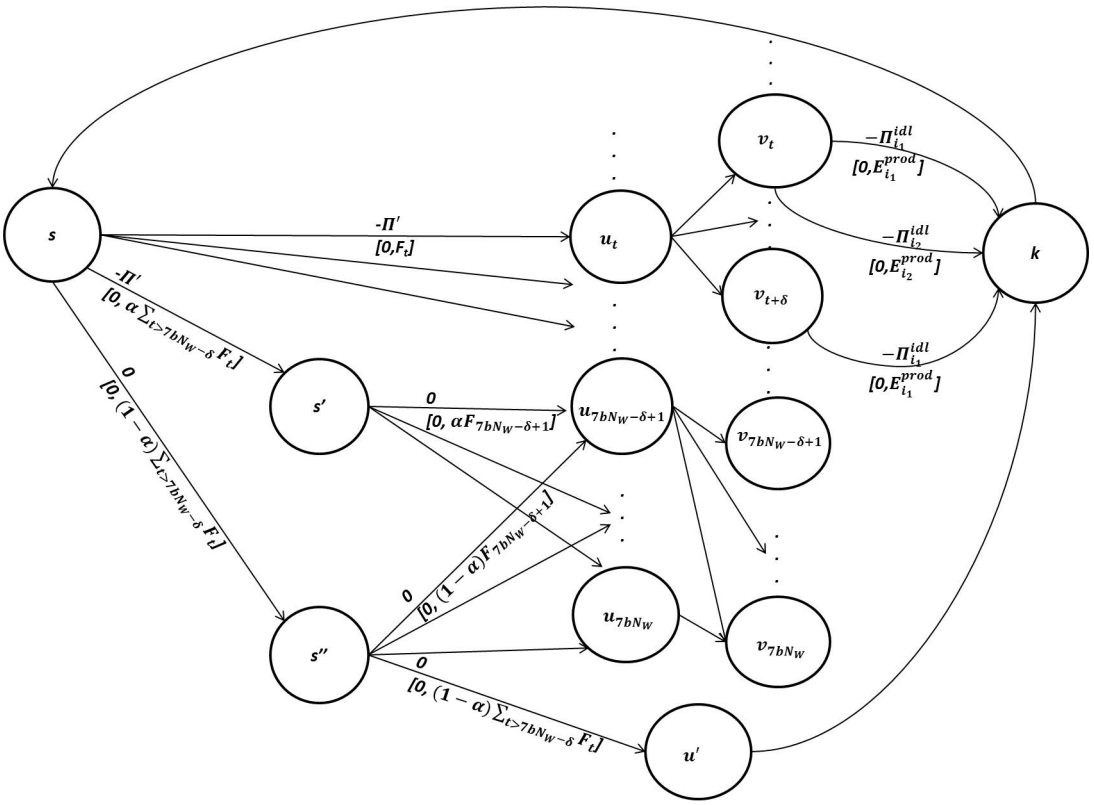


Figure 41: Network representation of the Fixed Schedule Model (FSM).

APPENDIX F

CHAPTER 3: WDSM IMPLEMENTATION

The following preferences and constraints are incorporated into WDSM based on the company's practices and policies:

- The schedule planning horizon is two weeks.
- The time window for demand fulfillment is 20 hours (5 time buckets).
- Given the potential choice of shifts, each day is divided into $b = 6$ time buckets of 4 hours each, with the first bucket starting at 7:00AM.
- Given a daily forecast (provided by its clients), the company creates a forecast for each time bucket.
- There is a minimum of 12-hour rest between shifts. All pairs of shifts that do not have a 12-hour rest between them are included in the set Q^{stro}. The schedule from the previous planning horizon is also considered to ensure that the rest-constraints are satisfied during the transition from one planning horizon to the next.
- Each employee i can be scheduled for a maximum number of hours per week ($E_{i,w}^{h_{max}}$), varying by employee type and productivity level. Similarly, there is a desired minimum number of hours per employee ($E_{i,w}^{h_{min}}$), which is also adjusted by employee type and productivity, and if the employee requested days off. This is modeled through the penalty $\Pi_i^{h_{min}}$.
- The maximum number of employees scheduled at any time is fixed in location l and equals the number of computers (or seats) available ($M_{t,l}$).

- An employee i might not be available on a given day or shift (for instance, some women request not to work during night shifts). This is captured in the sets O_i. If a given shift is not available at all, then all the staff need to be 'off' during that shift.

- Within the planning horizon of the schedule, it is preferred that all shifts for one employee are of the same type (morning, afternoon, or night). If a shift of a different type is assigned within a specified number of days, there is a penalty. These undesired shift-day pairs duples were modeled through the set Q^{soft} and the corresponding penalties $\Pi^{soft}_{i,(j_1,d_1),(j_2,d_2)}$, $(j_1, d_1), (j_2, d_2) \in Q^{soft}$.

- There is a preference to keep consistency across consecutive weeks within a schedule planning horizon (for instance, working the same shift on both Mondays). This preference is modeled by the penalty $\Pi^{dif}_{i,w}$.

- However, since the company also wants to rotate the staff across the different days and shifts, the preference penalties $\Pi^{pref}_{i,j,d}$ are computed such as there is a higher penalty if an employee i is assigned to the same types of shifts and days assigned during *previous* schedules.

- Productivity rates E^{prod}_i are computed based on historical records.

- The company prioritizes employees with higher productivity. In particular, highest-producing employees must be scheduled and workload assigned trying to minimize their idle time. We sort the employees by productivity level, and in I^{fix} we include the subset with higher productivity until all of the forecasted demand could be covered if they worked at 100% utilization. If an employee is not required in the new schedule, he/she can be sent to training, assigned to other activities, or take time-off. Therefore, the lower productivity employees should be only scheduled when necessary.

- It is desired that at least 25% of the demand forecast arriving during a schedule's last night shift is completed during the planning horizon (i.e., $\alpha = 0.25$).

The model implementation started in April and stabilized through feedback and calibration (i.e., setting penalty values) around the end of October. Hence, in the computational study presented in this chapter, we used the data provided for the last 9 two-week schedules in 2012.

APPENDIX G

CHAPTER 3: SUPPORTING TOOLS FOR WDSM IMPLEMENTATION

We implemented WDSM in GLPK, which is a solver freely available for both commercial and academic purposes. We use GLPK for solving FSM when testing any given schedule. Moreover, GLPK was used to generate the .mps files to solve WDSM by any commercial solver of choice (e.g., CPLEX, Gurobi). We also developed a set of tools (Staffing Tool (ST), Penalty Input (PI), and Output Read (OR)) to help to generate input and analyze results during the scheduling process (using MS Excel and Visual Basic).

The schedule generation process starts with getting the input for the new run in ST. The calculation of the penalties can be done using PI by defining some parameters according to the user's preferences and priorities. Input files for the WDSM implementation in GLPK are created with ST, and then the resulting .mps file is solved using an optimization solver. Results are read with OR and then analyzed with ST. The proposed heuristic can be ran in ST as well. Figure 42 shows the proposed steps to generate the schedule. We describe ST and PI in more detail in the following sections.

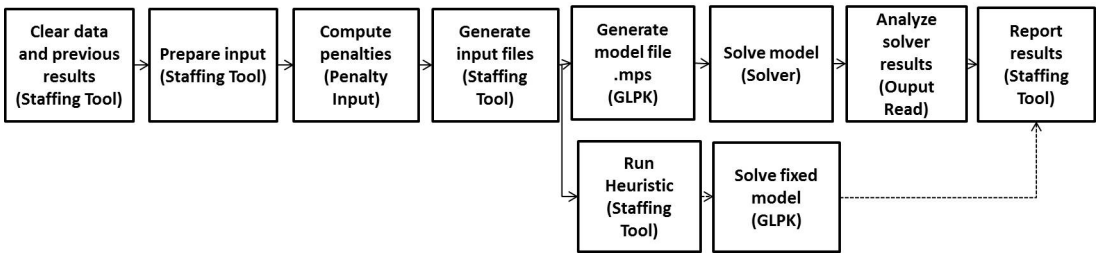


Figure 42: Proposed workflow to generate the next schedule.

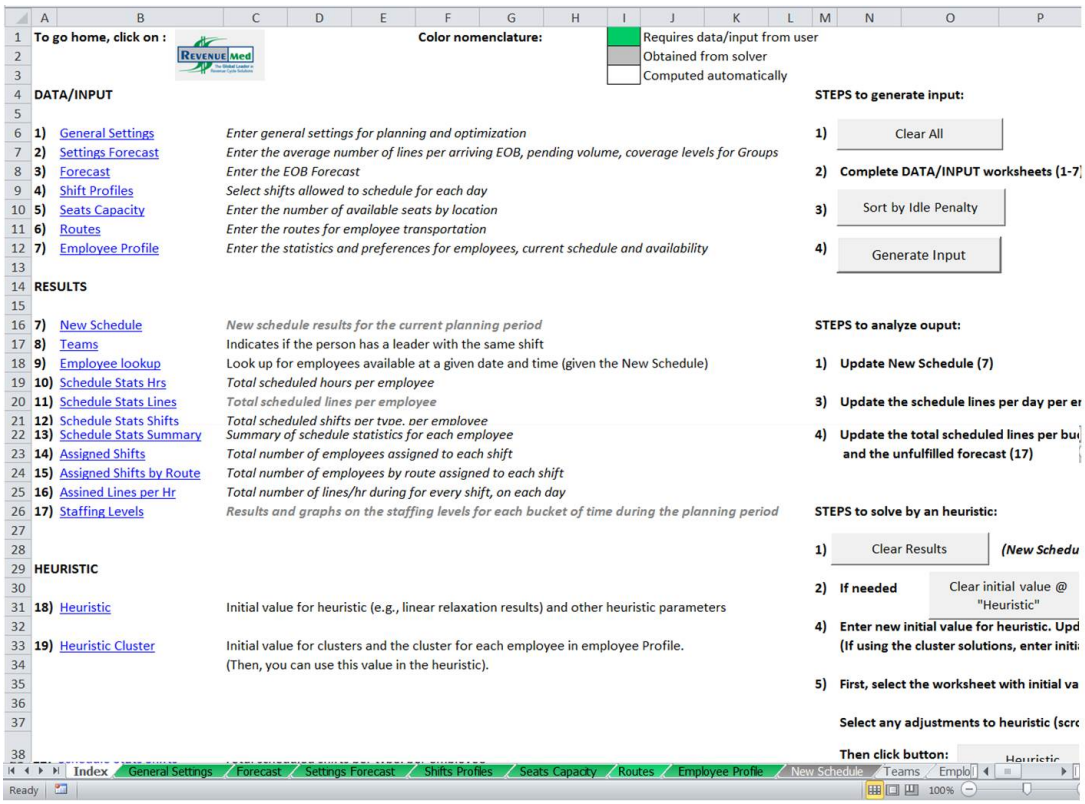


Figure 43: Snapshot of the Staff Tool (ST) main page.

G.1 Staffing Tool

ST is the main tool that contains all the input, results, and statistics. Also, it links to GLPK to generate both WDSM and FSM .mps files, and it runs the proposed heuristic described in Section 3.4.3. Figure 43 shows a snapshot of the ST main page (index).

Input worksheets:

- General settings: The user can modify general parameters, such as minimum/maximum hours per employee type, optimization settings, the demand fulfillment time window, penalties such as Π', etc.
- Forecast and Forecast settings.
- Shift profiles: Includes shift's start and end times and availability.

- Employee profile: Summarizes all the staff information including productivity, type, if they must be included in the schedule, preference penalties, schedule during the previous planning horizon, availability, etc.

Schedule statistics worksheets:

- Teams: This worksheet shows for every user, and for each assigned shift, if the team leader is also scheduled. Also, it shows the overall proportion of assigned shifts with leadership.

- Employee look-out: The user can select a date and time (from the planning horizon), and according to the schedule, find who may be available to be called to work if needed.

- Statistics, hours: Shows the number of hours scheduled to each employee and compares it to the week's minimum and maximum number of hours.

- Statistics, lines: Shows the workload allocated to each employee and compares it to the employee's productivity (as an estimated employee utilization).

- Statistics, shifts: Shows the type of shifts assigned to each employee. This includes performance metrics regarding the consistency and quality of the schedule.

- Staffing levels: Shows the overall statistics and graphs by time bucket, such as utilization of seat capacity, staff utilization, etc. The user can compare the forecasted demand, the processing capacity, the scheduled demand to fulfill, as well as the unfulfilled units of demand (see Figure 44), and analyze the effects of different demand scenarios.

Heuristics worksheets:

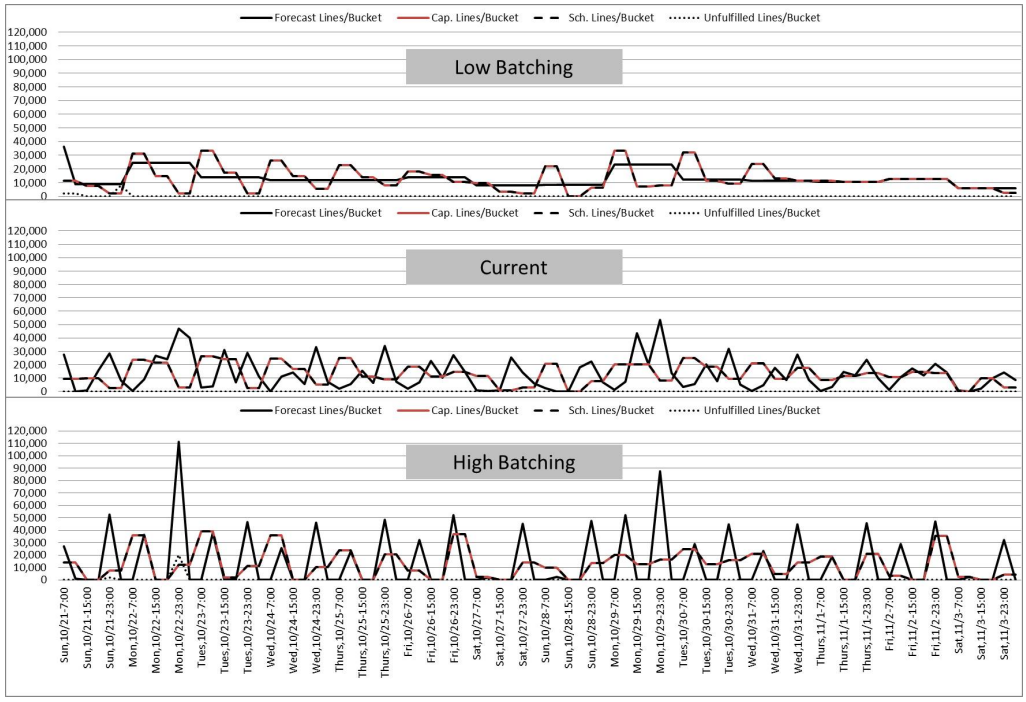


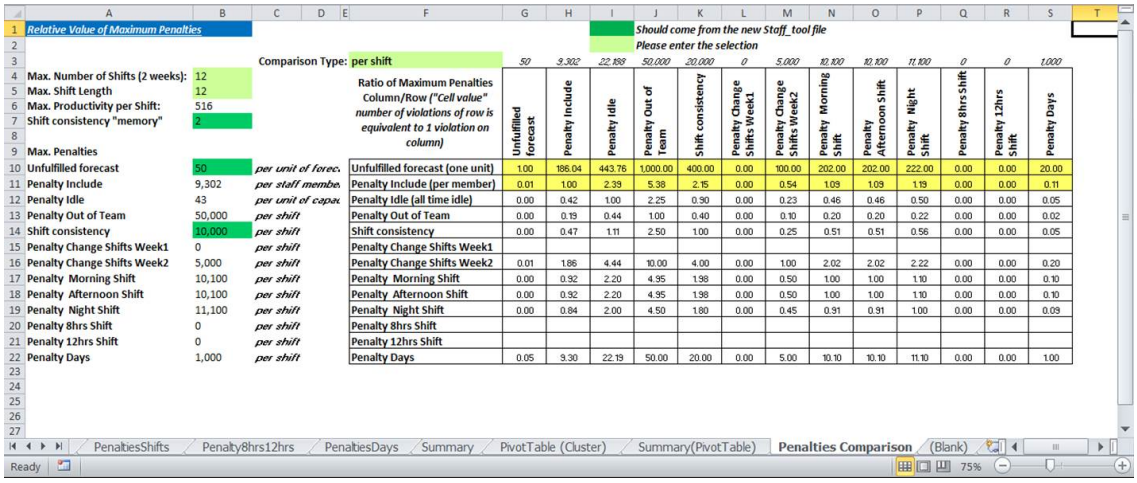
Figure 44: Example of a higher unfulfilled demand with a less variable demand flow.

- Heuristic: The user can enter the initial values or 'hints' for the heuristic and the additional penalty Π^{dev}.

- Heuristic (cluster): The user can specify the cluster each employee belongs to (if using cluster-based initial values).

G.2 Penalties Input

PI aids the user in calibrating the penalties. Most of the data comes directly from ST. Additionally, a history of up to two previous schedules (4 weeks) could be included to obtain a better rotation of the employees through different days of the week and in particular, different types of shifts (for instance, if an employee was assigned to only morning shifts before the previous two-week period, and only to nights shifts during the previous two-week period, then in the new schedule afternoons will be preferred). A shift ranking by employee can be also included. Once all the penalties are calculated in this file, a summary is created, which can be directly used in ST.



Relative Value of Maximum Penalties																							
		Comparison Type:	per shift			50	9,302	22,188	50,000	20,000	0	5,000	10,100	10,100	11,100	0	0	1,000					
Max. Number of Shifts (2 weeks):	12			Ratio of Maximum Penalties Column/Row ("Cell value" number of violations of row is equivalent to 1 violation on column)		Unfulfilled forecast	Penalty Include	Penalty Idle	Penalty Out of Team	Shift consistency	Penalty Change Shifts Week1	Penalty Change Shifts Week2	Penalty Morning Shift	Penalty Afternoon Shift	Penalty Night Shift	Penalty 8hrs Shift	Penalty 12hrs Shift	Penalty Days					
Max. Shift Length	12																						
Max. Productivity per Shift:	516																						
Shift consistency "memory"	2																						
Max. Penalties																							
Unfulfilled forecast	50	per unit of forec		Unfulfilled forecast (one unit)		1.00	186.04	443.76	1,000.00	400.00	0.00	100.00	202.00	202.00	222.00	0.00	0.00	20.00					
Penalty Include	9,302	per staff membe		Penalty Include (per member)		0.01	1.00	2.39	5.38	2.15	0.00	0.54	1.09	1.09	1.19	0.00	0.00	0.11					
Penalty Idle	43	per unit of capac		Penalty Idle (all time idle)		0.00	0.42	1.00	2.25	0.90	0.00	0.23	0.46	0.46	0.50	0.00	0.00	0.05					
Penalty Out of Team	50,000	per shift		Penalty Out of Team		0.00	0.19	0.44	1.00	0.40	0.00	0.10	0.20	0.20	0.22	0.00	0.00	0.02					
Shift consistency	10,000	per shift		Shift consistency		0.00	0.47	1.11	2.50	1.00	0.00	0.25	0.51	0.51	0.56	0.00	0.00	0.05					
Penalty Change Shifts Week1	0	per shift		Penalty Change Shifts Week1																			
Penalty Change Shifts Week2	5,000	per shift		Penalty Change Shifts Week2		0.01	1.86	4.44	10.00	4.00	0.00	1.00	2.02	2.02	2.22	0.00	0.00	0.20					
Penalty Morning Shift	10,100	per shift		Penalty Morning Shift		0.00	0.92	2.20	4.95	1.98	0.00	0.50	1.00	1.00	1.10	0.00	0.00	0.10					
Penalty Afternoon Shift	10,100	per shift		Penalty Afternoon Shift		0.00	0.92	2.20	4.95	1.98	0.00	0.50	1.00	1.00	1.10	0.00	0.00	0.10					
Penalty Night Shift	11,100	per shift		Penalty Night Shift		0.00	0.84	2.00	4.50	1.80	0.00	0.45	0.91	0.91	1.00	0.00	0.00	0.09					
Penalty 8hrs Shift	0	per shift		Penalty 8hrs Shift																			
Penalty 12hrs Shift	0	per shift		Penalty 12hrs Shift																			
Penalty Days	1,000	per shift		Penalty Days		0.05	9.30	22.19	50.00	20.00	0.00	5.00	10.10	10.10	11.10	0.00	0.00	1.00					

Should come from the new Staff_tool file

Please enter the selection

PenaltiesShifts | Penalty8hrs12hrs | PenaltiesDays | Summary | PivotTable (Cluster) | Summary(PivotTable) | Penalties Comparison | (Blank)

Ready

Figure 45: Matrix for penalties trade-off analysis.

The ratios among the penalties are computed to analyze the trade-offs between them (see Figure 45).

The penalties in PI take into account the number of people in the cluster when applicable. For instance, if a penalty for a shift is specified, it is multiplied by the number of people in the cluster to reflect the total penalty that would be incurred if the entire cluster is assigned to the given shift. PI can also aggregate the penalties of each employee by clusters and give a summary of penalties to be used in ST to model employee clusters.

APPENDIX H

CHAPTER 3: OUTLINE OF THE PROPOSED HEURISTIC ALGORITHM

Additional Sets/Lists, Parameters, and Variables

I^{sort}	Sorted list of employees $i \in I$, starting with team leaders $R \subseteq I$, followed by the rest of the employees, in order of decreasing productivity E_i^{prod}
Δ^{sort}	Sorted list of time bucket pairs $(t_1, t_2) \in \Delta$, in order of (i) increasing t_2 and (ii) t_1 with decreasing forecasted demand F_{t_1}
$ESP_i = \{_{(j,d):(j,d) \notin O_i, d \geq 1}\}$	Set of employee i's shift-day pairs (j, d) that are feasible to assign during the planning horizon
$VOL_{i,j,d,(t_1,t_2)}$	Units of forecasted demand arriving at time bucket t_1 that can be served by employee i during shift-day (j, d) and time bucket t_2
$NET^{\Pi}_{i,j,d}$	Net penalty resulting from scheduling employee i during shift-day (j, d), adjusted by any factor determined by the planner, minus the 'avoided' penalty for any served demand during the shift ($NET^{\Pi}_{i,j,d} < 0$ indicates a 'reward')
$HOUR_{i,w}$	Hours scheduled to employee i during week w

Initialize $HOUR_{i,w}$, $Y_{i,t}$ to zero
for $i \in I^{sort}$ **do**
  while $ESP_i \neq \emptyset$ **do**
    for $(j,d) \in ESP_i$ **do**
      if (j,d) is feasible to add to employee i's schedule **then**
        for $(t_1,t_2) \in \Delta^{sort}$ such that $(j,d) \in V_{t_2}$ **do**
          Calculate $VOL_{i,j,d,(t_1,t_2)}$
        end for
        Calculate $NET^{\Pi}_{i,j,d}$
      else
        $ESP_i \leftarrow ESP_i \setminus (j,d)$
      end if
    end for
    if $ESP_i \neq \emptyset$ **then**
      $(i,j^*,d^*) \leftarrow \arg\min_{ESP_i} NET^{\Pi}_{i,j,d}$
      $w^* \leftarrow \lceil d^*/7 \rceil$ (where $\lceil \cdot \rceil$ indicates to round up)
      if ($i \in I^{fix}$ and $\sum_{w \in W} HOUR_{i,w} = 0$) or ($\sum_{w \in W} HOUR_{i,w} > 0$ and $HOUR_{i,w^*} < E^{h_{min}}_{i,w^*}$) or ($NET^{\Pi}_{i,j^*,d^*} < 0$) **then**
        $X_{i,j^*,d^*} \leftarrow 1$
        $HOUR_{i,w^*} \leftarrow HOUR_{i,w^*} + H_{j^*}$
        for $(t_1,t_2) \in \Delta^{sort}$ such that $(j^*,d^*) \in V_{t_2}$ **do**
          $Y_{i,t_2} \leftarrow Y_{i,t_2} + VOL_{i,j^*,d^*,(t_1,t_2)}$
          $F_{t_1} \leftarrow F_{t_1} - VOL_{i,j^*,d^*,(t_1,t_2)}$
        end for
      end if
      $ESP_i \leftarrow ESP_i \setminus (j^*,d^*)$
    end if
  end while
end for

APPENDIX I

CHAPTER 4: COMPLEXITY PROOFS

We can model the general problem as a minimum cost flow problem with side constraints on a directed graph $G(V, E)$.

Graph G(V,E)

Vertices $v \in V$	
m_j, n_j	'Entry' and 'exit' vertices for jobs $j \in J$
s, t	'Source' and 'sink' vertices
Arcs $e \in E$	
(t, s)	Arcs between sink vertex t and source vertex s
(s, m_j)	Arcs between source vertex s and entry vertex m_j for jobs $j \in J$
(m_j, n_j)	Arcs between entry vertex m_j and exit vertex n_j for jobs $j \in J$
(n_j, t)	Arcs between exit vertex n_j and sink vertex t for jobs $j \in J$
(n_p, m_q)	Arcs between exit vertex n_p and entry vertex m_q such that $t_q \geq t_p + d_p$ for $p, q \in J$, $p \neq q$

The network flow variables $f_{e,i}$ indicate the quantity of resources type i circulating on arc e. A flow on arc (m_j, n_j) indicates a flow of resources going through job j, and a flow on arc (n_p, m_q) represents resources used in job p and used next in job q. A flow on arc (s, m_j) indicates that those resources were used for the first time for job j during the planning horizon, and similarly a flow on arc (n_j, t) indicates resources used for the last time for job j. Finally, as all the resources are collected in vertex t and start from vertex s, $f_{(t,s),i}$ represents the total quantity of resources type i circulating through the system.

An example of the described network $G(V, E)$ based on the jobs described in Figure 20 is shown in Figure 46. Job 1 ends before jobs 3, 4, 5, and 6 start, so it is connected with their entry vertices; similarly, jobs 2 and 4 end before jobs 5 and 6 start, and job 3 ends before jobs 4, 5, and 6 start. Source vertex s is connected to all jobs, and all jobs are connected to sink vertex t.

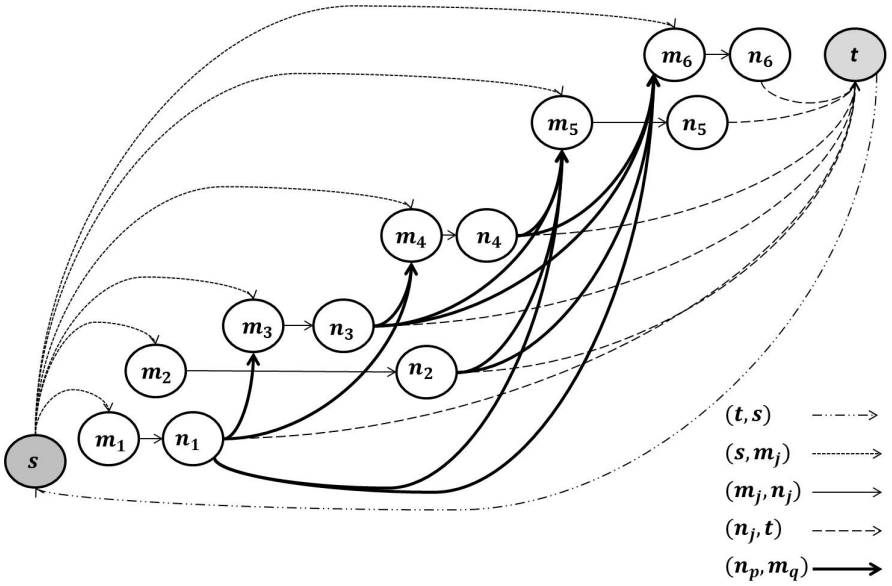


Figure 46: Example of network $G(V, E)$ with six jobs.

The resource planning and scheduling network model (RPSNM) is as follows:

$$MIN \quad Cost^{RPSNM} = \sum_{i \in I} c_i f_{(t,s),i} \tag{99}$$

$$\sum_{\substack{e \in E: \\ e=(v,p)}} f_{e,i} - \sum_{\substack{e \in E: \\ e=(p,v)}} f_{e,i} = 0 \quad i \in I,\ v \in V \tag{100}$$

$$f_{(m_j,n_j),i} \geq a_{i,j} x_j \quad i \in I,\ j \in J \tag{101}$$

$$\frac{1}{R} \sum_{j \in J} r_j x_j \geq \beta \tag{102}$$

$$\frac{1}{|J_k|} \sum_{j \in J_k} x_j \geq \beta_k \quad k \in K \tag{103}$$

$$f_{e,i} \in \mathbb{Z}_{\geq 0} \quad i \in I,\ e \in E \tag{104}$$

$$x_j \in \{0, 1\} \quad j \in J \tag{105}$$

Objective function (99) minimizes the total cost that results from the quantity of circulating resources. Constraints (100) are the network flow balance constraints. Constraints (101) set a lower bound that corresponds to the number of resources of type i required by job j if the job is accepted (i.e., $x_j = 1$), and zero otherwise. Constraint (102) and constraints (103) represent the global and 'per class' service requirements, respectively. Constraints (104) and (105) are non-negative integral and binary constraints of $f_{e,i}$ and x_j respectively.

The proofs of complexity for each of the special cases in **bold** in Figure 21 are next.

Proof. $\boldsymbol{K/I/a_{i,j}/1/0}$. It is easy to see that we can find the optimal inventory for each resource by tracking the number of resources required at each job's arrival, and finding the maximum number for each type across the planning horizon. In the example shown in Figure 20, we need two 'squares', four 'circles', and two 'triangles'. For a more formal proof, we can use the RPSNM formulation. If all the jobs are accepted, then $x_j = 1$ for all $j \in J$ and constraints (101) becomes:

$$f_{(m_j,n_j),i} \geq a_{i,j} \quad i \in I,\ j \in J \tag{106}$$

In addition we have that

$$\sum_{j \in J} r_j x_j = R$$

and

$$\sum_{j \in J_k} x_j = |J_k| \quad k \in K$$

and constraints (102) and (103) hold for any solution where all jobs are accepted, and therefore can be omitted. The problem is now solely described by constraints (100), (104) and (106), which together with objective function (99) correspond to a minimum cost circulation problem formulation, which is polynomially solvable. □

Proof. $\mathbf{1/I/a_{i,j}/\beta/0}$. Since there is only one class, all the jobs require the same resources. We treat all the resources needed for the only class as *one* bundle of resources (i.e., $|I| = 1$ and $a_{i,j} = 1$). Without loss of generality assume that the weight of each job of the only class and the cost of the resources bundle equal to 1. The network $G(V, E)$ is as described above. We assign a cost of -1 and an upper bound of 1 to arcs (m_j, n_j). We also set an upper bound of C to arc (t, s), which limits the number of resources (capacity) available in the system. Let N^{min} be the minimum *integer* number of jobs to accept, i.e., $\lceil R \cdot \beta \rceil = \lceil |J| \cdot \beta \rceil = N^{min}$. We can answer the question: can at least N^{min} jobs be accepted with C resources?, by solving a minimum cost circulation problem. If the answer to this question is 'no' we increase C, and we decrease C otherwise. We can find in polynomial time the minimum number of resources needed to accept at least N^{min} jobs by bisection on C. Finally, if there is a positive flow on arc (m_j, n_j), then job j is accepted. □

Proof. $\mathbf{K/1/1/\beta/0}$. This proof is very similar than the one for $1/I/a_{i,j}/\beta/0$, except that in this case there can be more than one demand class (i.e., $|K| \geq 1$). We assign a cost of $-r_j$ to arcs (m_j, n_j) and an upper bound of 1. We set an upper bound of C to arc (t, s). Let R^{min} be the minimum *integer* weighted jobs to accept, i.e., $\lceil R \cdot \beta \rceil = R^{min}$. We can answer the question: Can R^{min} weighted jobs be accepted with C resources?, by solving a minimum cost circulation problem. We can find in polynomial time the minimum number of resources needed to accept at least R^{min} weighted jobs by bisection on C. Also, job j is accepted if there is a positive flow on arc (m_j, n_j). □

Proof. $\mathbf{K/1/a_{i,j}/\beta/0}$. Recall the knapsack problem:

$$R = MAX \sum_{j \in J} r_j x_j$$

$$s.t. \sum_{j \in J} a_j x_j \leq C$$

$$x_j \in \{0,1\} \;\; j \in J$$

This instance of the knapsack problem is equivalent to the following instance of the special case $K/1/a_{i,j}/\beta/0$. There are J jobs, each one of a different class (i.e., $|K| = |J|$), and only one type of resource ($|I| = 1$); hence, $a_{i,j} = a_j$. All jobs are scheduled to start at the same time, have the same duration, and have weight r_j. By solving this problem, we can find the maximum weighted jobs $R = \sum_{j \in J} r_j x_j$ that require no more than C resource units to be accepted.

□

Proof. $\boldsymbol{K/I/ \leq 1/\beta/0}$. We will reduce the exact cover by 3-sets (X3C) problem to an instance of special case $K/I/ \leq 1/\beta/0$. In the X3C, there is finite set I with $3q$ unique elements and a collection Z of $p \geq q$ 3-element subsets of I, $Z = \{Z_1, \ldots, Z_p\}$. To solve X3C we answer the question: does Z contain an exact cover for I, that is, a subcollection $Z^* \subseteq Z$ such that every element of I occurs in exactly one member of Z^*?

Given an instance of the X3C, we create an instance of problem $K/I/ \leq 1/\beta/0$. Each element in I corresponds to a type of resource and $|I| = 3q$. There are a total of $|K| = p + 3q$ demand classes, one for each subset in Z and one for each element in I. There is one job per class (i.e., $|J| = |K|$). For each subset in Z, there is a job scheduled to start at time zero with a duration of $d/2$ ($j = 1, \ldots, p$). Each of these p jobs require one unit of the three types of resources that correspond to the elements of the subset in Z. Also, there are $3q$ jobs that start at $d/2$ with a duration of $d/2$ ($j = p+1, \ldots, p+3q$). Each of these $3q$ jobs require one of unit of a different resource in I. We assume that $r_j = 1$ for all jobs and that $\beta \geq \frac{4q}{p+3q}$, which means that at least $4q$ jobs should be accepted. We have that $c_i = 1$ for all types of resources and that the total cost of the available resources is $3q$, i.e, there are $3q$ resources available. Since there are only $3q$ resources available at time zero, and each of the p jobs starting at that time require exactly 3 resources, a maximum of q jobs can be accepted from

time zero to time $d/2$. Then, there are exactly $3q$ jobs starting at $d/2$, and all of the $3q$ jobs must be accepted to be able to fulfill the minimum $4q$ jobs. This means that for a feasible solution where the total cost is $3q$, there cannot be any idle time for the available resources. It also implies that there is one resource unit on inventory of each one of the $3q$ different resource types.

It is easy to see that if there is a solution for this instance of problem $K/I/ \leq 1/\beta/0$, there is a solution for the original X3C problem. The accepted q jobs starting at time zero and ending at time $d/2$ correspond to the q subsets of Z that make an exact cover by 3-sets of the set I. They are an exact cover since all the $3q$ resources used for these q jobs are used next on the $3q$ jobs starting at $d/2$, each requiring one different resource, covering each one of the $3q$ elements in I.

□

Proof. $\boldsymbol{K/1/1/0/\beta_k}$. We will reduce the numerical 3-dimensional matching problems (N3DM) to an instance of special case $K/1/1/0/\beta_k$. In the N3DM, there are integers C, H, and g_p, y_p, z_p for $p = 1, \ldots, C$ satisfying the following:

$$\sum_{p=1}^{C} (g_p + y_p + z_p) = C \cdot H$$

and

$$0 < g_p, y_p, z_p < H \quad p = 1, \ldots, C$$

The objective is to find λ and δ of $\{1, \ldots, C\}$, such that:

$$g_p + y_{\lambda(p)} + z_{\delta(p)} = H \quad p = 1, \ldots, C$$

In what follows on this proof, we represent each job j by its scheduled start time t_j, scheduled end time $t_j + d_j$, and the job's class $k_j \in K$ as $\left(t_j, t_j + d_j\right)_{k_j}$. We create an instance of problem $K/1/1/0/\beta_k$ as follows:

Every job requires one unit of the only resource available ($|I| = 1$). We generate a first group of C jobs ($j = 1, \ldots, C$). Job p starts at time zero, with a duration g_p, and class p, for $p = 1, \ldots, C$:

$$\big(0, g_p\big)_p \quad p = 1, \ldots, C$$

Then, we generate a second group of C^2 jobs ($j = C+1, \ldots, C^2+C$). Job $qC+p$ starts at time g_p, with a duration y_q, and class $C+q$, for $p, q = 1, \ldots, C$:

$$\big(g_p, g_p + y_q\big)_{C+q} \quad p, q = 1, \ldots, C$$

Finally, we create a third group of C jobs ($j = C^2 + C + 1, \ldots, C^2 + 2C$). Job $C^2 + C + p$ starts on time $H - z_p$ (where H is the schedule planning horizon), with a duration of z_p, and class $2C + p$, for $p = 1, \ldots, C$:

$$\big(H - z_p, H\big)_{2C+p} \quad p = 1, \ldots, C$$

We set a capacity (cost) of C resources, and the minimum number of jobs to be accepted should be at least one for each class $k \in K$ (where $|K| = 3C$). Note that there is exactly one job of each class $k = 1, \ldots, C$, so each one must be accepted. Therefore, the sum of these jobs' durations (first group) is $\sum_{p=1}^{C} g_p$. Similarly, there is exactly one job for each class $k = 2C+1, \ldots, 3C$, and each one must be accepted. The sum of these jobs' durations (third group) is $\sum_{p=1}^{C} z_p$. Finally, there are C jobs for *each* class $k = C + 1, \ldots, 2C$, and at least one job of each class should be accepted. Then, the sum of these accepted jobs durations (second group) should be at least $\sum_{p=1}^{C} y_p$. As a result, the sum of all the accepted jobs' durations should be at least:

$$\sum_{p=1}^{C} g_p + \sum_{p=1}^{C} y_p + \sum_{p=1}^{C} z_p = \sum_{p=1}^{C} (g_p + y_p + z_p) = C \cdot H$$

However, since all the C jobs of class $k = 1, \ldots, C$ (first group) start at time zero, and all the C jobs of class $k = 2C + 1, \ldots, 3C$ (third group) end at time H,

and all these jobs should be accepted, then all the C resources available should start their schedule at time zero and end at time H. It follows that the maximum sum of durations of all the accepted jobs can be at most $C \cdot H$ (with no idle time), which is also the minimum sum of durations of the accepted jobs. This means that if there is a solution with a capacity of C resources, then there is no idle time, i.e, all the C resources are used during the whole time interval $[0, H]$. It follows that the C resources must have schedules of the form:

$$\big(0, g_p\big)_p \big(g_p, g_p + y_q\big)_{C+q} \big(H - z_r, H\big)_{2C+r}$$

From these schedules and the fact that there is not idle time, we get that we must have $g_p + y_q = H - z_r$, i.e., $g_p + y_q + z_r = H$. It is easy to map the solution to this problem $K/1/1/0/\beta_k$ to the original N3DM problem: we define $\gamma(p) = q$ and $\delta(p) = r$ given the schedules of each of the C resources, starting with job of class p, for $p = 1, \ldots, C$.

□

Finally, since problem $K/I/a_{i,j}/1/0$ is polynomially solvable, it follows that all cases where all jobs should be accepted are polynomially solvable (Figure 21, first column). Also, since problems $K/1/a_{i,j}/\beta/0$, $K/I/ \leq 1/\beta/0$ and $K/1/1/0/\beta_k$ are NP-complete, it follows that cases $K/I/a_{i,j}/\beta/0$, $K/1/a_{i,j}/0/\beta_k$, $K/I/ \leq 1/0/\beta_k$, and $K/I/a_{i,j}/0/\beta_k$ are NP-complete as the former are special cases of the latter (see Figure 21).

APPENDIX J

CHAPTER 4: SUMMARY OF RELEVANT RESULTS BY WANG AND AHMED

The main technique used to derive the convergence of the SAA method is Large Deviations (LD) theory. Consider a random variable Y, with mean $\mu = \mathbb{E}[Y]$. Let $\gamma(u) = \sup_{s\in\mathbb{R}}\{su - \Lambda(s)\}$ for $u \in \mathbb{R}$, where $\Lambda(s) = \log(G(s))$, and $G(s) = \mathbb{E}[e^{sY}]$ is the moment generating function (MGF) of Y. Consider an i.i.d sequence of $Y_1, \ldots, Y_M$ replications of Y, and let $y^M = \frac{1}{M}\sum_{m=1}^{M} Y_m$ be the sample average. Then, LD theory states that for any real number $v > \mu$,

$$P(y^M \geq v) \leq e^{-M\gamma(v)} \tag{107}$$

Similarly, if $v < \mu$,

$$P(y^M \leq v) \leq e^{-M\gamma(v)} \tag{108}$$

Moreover, if the MGF $G(s)$ is finite for all s in a neighborhood of $s = 0$, by Taylor's expansion,

$$\gamma(v) = \frac{(v-\mu)^2}{2Var[Y]} + o(|v-\mu|^2) \tag{109}$$

Equation (109) implies $\gamma > 0$. Also, for v close enough to μ, $\gamma(v)$ can be approximated by $\frac{(v-\mu)^2}{2Var[Y]}$.

Wang and Ahmed [122] consider an stochastic problem of the form:

$$\min_{z\in Z}\{g(z) : h(z) := \mathbb{E}[H(z,\xi)] \leq b\}$$

Z is the set of feasible decisions, ξ is a random vector with support Ω, $g : Z \to \mathbb{R}$, and $H : Z \times \Omega \to \mathbb{R}$. Given a sample $\{\xi_1, \ldots, \xi_M\}$, the SAA of the problem is given by:

$$\min_{z \in Z} \left\{ g(z) : h^M(z) := \frac{1}{M} \sum_{m=1}^{M} H(z, \xi) \leq b \right\}$$

Given an $\epsilon > 0$, define the solution regions $Z^{+\epsilon}$ and $Z^{-\epsilon}$ of the original problem as:

$$Z^{+\epsilon} = \{z \in Z : h(z) \leq b + \epsilon\}, \;\; Z^{-\epsilon} = \{z \in Z : \; h(z) \leq b - \epsilon\}$$

Also, define the feasible solution region for the SAA as:

$$Z^M = \{z \in Z : h^M(z) \leq b\}$$

Suppose conditions C1-C3 hold:

(C1) Z is a nonempty compact set.

(C2) h(z) is well-defined, i.e., for every $z \in Z$, $H(z, \cdot)$ is measurable and $\mathbb{E}|H(z, \cdot)| < +\infty$

(C3) For any $z \in Z$ the MGF of $H(z, \xi) - h(z)$ is finite in a neighborhood of zero.

In addition, if $|Z|$ is finite, the following result follows:

$$P(Z^{-\epsilon} \subseteq Z^M \subseteq Z^{+\epsilon}) \geq 1 - 2|Z|e^{-\frac{M\epsilon^2}{2\sigma^2}}$$

where $\sigma^2 = \max_{z \in Z} Var[H(z, \xi) - h(z)]$.

APPENDIX K

CHAPTER 4: SAARPM CONVERGENCE PROOF FOR THE CASE WITH A SINGLE EXPECTATION CONSTRAINT

Under the compactness assumption, since F is bounded with integral coordinates, it follows that $|F|$ is finite. By the definitions for F, $F^{+\epsilon}$, $F^{-\epsilon}$, and $F^{|S|}$,

$$P(F^{+\epsilon} \subseteq F^{|S|} \subseteq F^{-\epsilon}) = 1 - P(\exists \mathbf{f} \in F : \ \mathbf{f} \in F^{+\epsilon} \, and \, \mathbf{f} \notin F^{|S|}, \, or \, \mathbf{f} \in F^{|S|} \, and \, \mathbf{f} \notin F^{-\epsilon})$$

$l^{max}(\mathbf{f}) \geq \beta + \epsilon$ ($l^{max}(\mathbf{f}) < \beta + \epsilon$) if and only if $\mathbf{f} \in F^{+\epsilon}$ ($\mathbf{f} \notin F^{+\epsilon}$). The same relationship applies between $l^{max}(\mathbf{f})$, $\beta - \epsilon$, and $F^{-\epsilon}$; and $l^{max_{|S|}}(\mathbf{f})$, β, and $F^{|S|}$. We follow the proof presented in [122],

$$\begin{aligned}
P(F^{+\epsilon} \subseteq F^{|S|} \subseteq F^{-\epsilon}) &= 1 - P(\exists \mathbf{f} \in F : \ l^{max}(\mathbf{f}) \geq \beta + \epsilon \, and \, l^{max_{|S|}}(\mathbf{f}) < \beta, \\
&\qquad or \, l^{max_{|S|}}(\mathbf{f}) \geq \beta \, and \, l^{max}(\mathbf{f}) < \beta - \epsilon) \\
&\geq 1 - P(\exists \mathbf{f} \in F : \ l^{max}(\mathbf{f}) \geq \beta + \epsilon \, and \, l^{max_{|S|}}(\mathbf{f}) < \beta) \\
&\quad - P(\exists \mathbf{f} \in F : \ l^{max_{|S|}}(\mathbf{f}) \geq \beta \, and \, l^{max}(\mathbf{f}) < \beta - \epsilon) \\
&\geq 1 - P(\exists \mathbf{f} \in F : \ l^{max_{|S|}}(\mathbf{f}) - l^{max}(\mathbf{f}) < -\epsilon) \\
&\quad - P(\exists \mathbf{f} \in F : \ l^{max_{|S|}}(\mathbf{f}) - l^{max}(\mathbf{f}) > \epsilon) \\
&\geq 1 - P(\exists \mathbf{f} \in F : \ l^{max_{|S|}}(\mathbf{f}) - l^{max}(\mathbf{f}) \leq -\epsilon) \\
&\quad - P(\exists \mathbf{f} \in F : \ l^{max_{|S|}}(\mathbf{f}) - l^{max}(\mathbf{f}) \geq \epsilon) \\
&\geq 1 - \sum_{\mathbf{f} \in F} P(l^{max_{|S|}}(\mathbf{f}) - l^{max}(\mathbf{f}) \leq -\epsilon) - \sum_{\mathbf{f} \in F} P(l^{max_{|S|}}(\mathbf{f}) - l^{max}(\mathbf{f}) \geq \epsilon)
\end{aligned}$$

By LD inequalities (107) and (108), $P(l^{max_{|S|}}(\mathbf{f}) - l^{max}(\mathbf{f}) \leq -\epsilon) \leq e^{-|S|\gamma_{\mathbf{f}}^{Lmax}(-\epsilon)}$ and $P(l^{max_{|S|}}(\mathbf{f}) - l^{max}(\mathbf{f}) \geq \epsilon) \leq e^{-|S|\gamma_{\mathbf{f}}^{Lmax}(\epsilon)}$. Then,

$$P(F^{+\epsilon} \subseteq F^{|S|} \subseteq F^{-\epsilon}) \geq 1 - \sum_{\mathbf{f} \in F} e^{-|S|\gamma_{\mathbf{f}}^{Lmax}(-\epsilon)} - \sum_{\mathbf{f} \in F} e^{-|S|\gamma_{\mathbf{f}}^{Lmax}(\epsilon)}$$

Then, since the MGF of $L^{max}(\mathbf{f}, w) - l^{max}(\mathbf{f})$ is finite for all $\mathbf{f} \in F$, by equation (109),

$$\gamma_{\mathbf{f}}^{Lmax}(-\epsilon), \gamma_{\mathbf{f}}^{Lmax}(\epsilon) \geq \frac{\epsilon^2}{2Var[L^{max}(\mathbf{f}, w) - l^{max}(\mathbf{f})]}$$

Then,

$$P(F^{+\epsilon} \subseteq F^{|S|} \subseteq F^{-\epsilon}) \geq 1 - 2\sum_{\mathbf{f} \in F} e^{-|S|\frac{\epsilon^2}{2Var[L^{max}(\mathbf{f}, w) - l^{max}(\mathbf{f})]}}$$

Following the definition of σ_{Lmax}^2,

$$P(F^{+\epsilon} \subseteq F^{|S|} \subseteq F^{-\epsilon}) \geq 1 - 2\sum_{\mathbf{f} \in F} e^{-\frac{|S|\epsilon^2}{2\sigma_{Lmax}^2}}$$

Finally, taking the summation over $\mathbf{f} \in F$,

$$P(F^{+\epsilon} \subseteq F^{|S|} \subseteq F^{-\epsilon}) \geq 1 - 2|F|e^{-\frac{|S|\epsilon^2}{2\sigma_{Lmax}^2}}$$

APPENDIX L

CHAPTER 4: SAARPM CONVERGENCE PROOF FOR THE CASE WITH MULTIPLE EXPECTATION CONSTRAINTS

The proof is an extension of Proposition 4.5.1's. F is bounded with integral coordinates, therefore $|F|$ is finite. By the definitions for F, $F^{+\Delta}$, $F^{-\Delta}$, and $F^{|S|}$,

$$\begin{aligned} P(F^{+\Delta} \subseteq F^{|S|} \subseteq F^{-\Delta}) &= 1 - P(\exists \mathbf{f} \in F \; s.t. \; \mathbf{f} \in F^{+\Delta} \; and \; \mathbf{f} \notin F^{|S|}, \\ &\qquad or \; \mathbf{f} \in F^{|S|} \; and \; \mathbf{f} \notin F^{-\Delta}) \\ &\geq 1 - P(\exists \mathbf{f} \in F \; s.t. \; \mathbf{f} \in F^{+\Delta} \; and \; \mathbf{f} \notin F^{|S|}) \\ &\qquad - P(\exists \mathbf{f} \in F \; s.t. \; \mathbf{f} \in F^{|S|} \; and \; \mathbf{f} \notin F^{-\Delta}) \end{aligned}$$

First consider the case $\exists \mathbf{f} \in F$ s.t. $\mathbf{f} \in F^{+\Delta}$ and $\mathbf{f} \notin F^{|S|}$: Since there exist a vector $\mathbf{f} \in F$ such that $\mathbf{f} \in F^{+\Delta}$, then there is at least one vector for the accept/reject decisions, say $\mathbf{x}^{+\Delta} \in X(\mathbf{f})$, such that constraints (83) and (84) hold. This implies that given $\mathbf{x}^{+\Delta}$, the expected GWSL, $l(\mathbf{f}, \mathbf{x}^{+\Delta})$, is greater or equal than $\beta + \epsilon$; and that expected CSL, $l_k(\mathbf{f}, \mathbf{x}^{+\Delta})$, is greater or equal than $\beta_k + \epsilon_k$. Since $\mathbf{f} \notin F^{|S|}$, *one* or more service constraints are violated in SAAPRM given $\mathbf{f}$ resources available. Consider again the vector of accept/reject decisions $\mathbf{x}^{+\Delta} \in X(\mathbf{f})$. Given capacity decisions $\mathbf{f} \in F$, the probability of constraint (80) to be violated is not greater than the probability of $l^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) < \beta$, where jobs' accept/reject decisions are fixed. Similarly, the probability of a constraint (81) for class k to be violated is not greater than the probability of $l_k^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) < \beta_k$. Note that $l_k^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta})$ is defined for all $\mathbf{f} \in F$ and all $k \in K$ since the sample S is class-representative. Therefore,

$$P(F^{+\Delta} \subseteq F^{|S|} \subseteq F^{-\Delta}) \geq 1 - P(\exists \mathbf{f} \in F \; s.t. \; l(\mathbf{f}, \mathbf{x}^{+\Delta}) \geq \beta + \epsilon \; and \; l^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) < \beta)$$
$$- \sum_{k \in K} P(\exists \mathbf{f} \in F \; s.t. \; l_k(\mathbf{f}, \mathbf{x}^{+\Delta}) \geq \beta_k + \epsilon_k \; and \; l_k^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) < \beta_k)$$
$$- P(\exists \mathbf{f} \in F \; s.t. \; \mathbf{f} \in F^{|S|} \; and \; \mathbf{f} \notin F^{-\Delta})$$

No consider $\exists \mathbf{f} \in F$ s.t. $\mathbf{f} \in F^{|S|}$ and $\mathbf{f} \notin F^{-\Delta}$: Since there exist a vector $\mathbf{f} \in F$ such that $\mathbf{f} \in F^{|S|}$, then there is at least one vector for the accept/reject decisions, say $\mathbf{x}^{|S|} \in X(\mathbf{f})$, such that constraints (80) and (81) hold. This implies that the sample average GWSL given $\mathbf{x}^{|S|}$, $l^{|S|}(\mathbf{f}, \mathbf{x}^{|S|})$, is greater or equal than β; and that the sample average CSL, $l_k^{|S|}(\mathbf{f}, \mathbf{x}^{|S|})$, is greater or equal than β_k. Since $\mathbf{f} \notin F^{-\Delta}$, *one* or more service constraints are violated in SPRM($-\Delta$). Consider the accept/reject decisions given by $\mathbf{x}^{|S|} \in X(\mathbf{f})$. The probability of constraint (85) to be violated is not greater than the probability of $l(\mathbf{f}, \mathbf{x}^{|S|}) < \beta - \epsilon$, where jobs' accept/reject decisions are fixed to $\mathbf{x}^{|S|}$. Similarly, the probability of a constraint (86) for class $k \in K$ to be violated is not greater than the probability of $l_k(\mathbf{f}, \mathbf{x}^{|S|}) < \beta_k - \epsilon_k$.

$$
\begin{aligned}
P(F^{+\Delta} \subseteq F^{|S|} \subseteq F^{-\Delta}) \geq 1 &- P(\exists \mathbf{f} \in F \; s.t. \; l(\mathbf{f}, \mathbf{x}^{+\Delta}) \geq \beta + \epsilon \; and \; l^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) < \beta) \\
&- \sum_{k \in K} P(\exists \mathbf{f} \in F \; s.t. \; l_k(\mathbf{f}, \mathbf{x}^{+\Delta}) \geq \beta_k + \epsilon_k \; and \; l_k^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) < \beta_k) \\
&- P(\exists \mathbf{f} \in F \; s.t. \; l^{|S|}(\mathbf{f}, \mathbf{x}^{|S|}) \geq \beta \; and \; l(\mathbf{f}, \mathbf{x}^{|S|}) < \beta - \epsilon) \\
&- \sum_{k \in K} P(\exists \mathbf{f} \in F \; s.t. \; l_k^{|S|}(\mathbf{f}, \mathbf{x}^{|S|}) \geq \beta_k \; and \; l_k(\mathbf{f}, \mathbf{x}^{|S|}) < \beta_k - \epsilon_k) \\
\geq 1 &- P(\exists \mathbf{f} \in F \; s.t. \; l^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) - l(\mathbf{f}, \mathbf{x}^{+\Delta}) < -\epsilon) \\
&- \sum_{k \in K} P(\exists \mathbf{f} \in F \; s.t. \; l_k^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) - l_k(\mathbf{f}, \mathbf{x}^{+\Delta}) < -\epsilon_k) \\
&- P(\exists \mathbf{f} \in F \; s.t. \; l^{|S|}(\mathbf{f}, \mathbf{x}^{|S|}) - l(\mathbf{f}, \mathbf{x}^{|S|}) > \epsilon) \\
&- \sum_{k \in K} P(\exists \mathbf{f} \in F \; s.t. \; l_k^{|S|}(\mathbf{f}, \mathbf{x}^{|S|}) - l_k(\mathbf{f}, \mathbf{x}^{|S|}) > \epsilon_k) \\
\geq 1 &- P(\exists \mathbf{f} \in F \; s.t. \; l^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) - l(\mathbf{f}, \mathbf{x}^{+\Delta}) \leq -\epsilon) \\
&- \sum_{k \in K} P(\exists \mathbf{f} \in F \; s.t. \; l_k^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) - l_k(\mathbf{f}, \mathbf{x}^{+\Delta}) \leq -\epsilon_k) \\
&- P(\exists \mathbf{f} \in F \; s.t. \; l^{|S|}(\mathbf{f}, \mathbf{x}^{|S|}) - l(\mathbf{f}, \mathbf{x}^{|S|}) \geq \epsilon) \\
&- \sum_{k \in K} P(\exists \mathbf{f} \in F \; s.t. \; l_k^{|S|}(\mathbf{f}, \mathbf{x}^{|S|}) - l_k(\mathbf{f}, \mathbf{x}^{|S|}) \geq \epsilon_k) \\
\geq 1 &- \sum_{\mathbf{f} \in F} P(l^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) - l(\mathbf{f}, \mathbf{x}^{+\Delta}) \leq -\epsilon) \\
&- \sum_{k \in K} \sum_{\mathbf{f} \in F} P(l_k^{|S|}(\mathbf{f}, \mathbf{x}^{+\Delta}) - l_k(\mathbf{f}, \mathbf{x}^{+\Delta}) \leq -\epsilon_k) \\
&- \sum_{\mathbf{f} \in F} P(l^{|S|}(\mathbf{f}, \mathbf{x}^{|S|}) - l(\mathbf{f}, \mathbf{x}^{|S|}) \geq \epsilon) \\
&- \sum_{k \in K} \sum_{\mathbf{f} \in F} P(l_k^{|S|}(\mathbf{f}, \mathbf{x}^{|S|}) - l_k(\mathbf{f}, \mathbf{x}^{|S|}) \geq \epsilon_k)
\end{aligned}
$$

By LD theory inequalities (107) and (108),

$$P(F^{+\Delta} \subseteq F^{|S|} \subseteq F^{-\Delta}) \geq 1 - \sum_{\mathbf{f} \in F} e^{-|S|\gamma_{\mathbf{f},\mathbf{x}^{+\Delta}}(-\epsilon)} - \sum_{k \in K} \sum_{\mathbf{f} \in F} e^{-|S|\gamma_{k,\mathbf{f},\mathbf{x}^{+\Delta}}(-\epsilon_k)} - \sum_{\mathbf{f} \in F} e^{-|S|\gamma_{\mathbf{f},\mathbf{x}^{|S|}}(\epsilon)} - \sum_{k \in K} \sum_{\mathbf{f} \in F} e^{-|S|\gamma_{k,\mathbf{f},\mathbf{x}^{|S|}}(\epsilon_k)}$$

Then, since the MGF of $L(\mathbf{f}, \mathbf{x}, w) - l(\mathbf{f}, \mathbf{x})$ and $L_k(\mathbf{f}, \mathbf{x}, w) - l_k(\mathbf{f}, \mathbf{x})$ are finite for all $\mathbf{x} \in X(\mathbf{f})$ and $\mathbf{f} \in F$, by equation (109),

$$\gamma_{\mathbf{f},\mathbf{x}^{+\Delta}}(-\epsilon) \geq \frac{\epsilon^2}{2Var[L(\mathbf{f}, \mathbf{x}^{+\Delta}, w) - l(\mathbf{f}, \mathbf{x}^{+\Delta})]},$$
$$\gamma_{\mathbf{f},\mathbf{x}^{|S|}}(\epsilon) \geq \frac{\epsilon^2}{2Var[L(\mathbf{f}, \mathbf{x}^{|S|}, w) - l(\mathbf{f}, \mathbf{x}^{|S|})]}$$

and,

$$\gamma_{k,\mathbf{f},\mathbf{x}^{+\Delta}}(-\epsilon_k) \geq \frac{\epsilon_k^2}{2Var[L_k(\mathbf{f}, \mathbf{x}^{+\Delta}, w) - l_k(\mathbf{f}, \mathbf{x}^{+\Delta})]},$$
$$\gamma_{k,\mathbf{f},\mathbf{x}^{|S|}}(\epsilon_k) \geq \frac{\epsilon_k^2}{2Var[L_k(\mathbf{f}, \mathbf{x}^{|S|}, w) - l_k(\mathbf{f}, \mathbf{x}^{|S|})]} \quad k \in K$$

Then,

$$P(F^{+\Delta} \subseteq F^{|S|} \subseteq F^{-\Delta}) \geq 1 - \sum_{\mathbf{f} \in F} e^{-\frac{|S|\epsilon^2}{2Var[L(\mathbf{f},\mathbf{x}^{+\Delta},w) - l(\mathbf{f},\mathbf{x}^{+\Delta})]}} - \sum_{k \in K} \sum_{\mathbf{f} \in F} e^{-\frac{|S|\epsilon_k^2}{2Var[L_k(\mathbf{f},\mathbf{x}^{+\Delta},w) - l_k(\mathbf{f},\mathbf{x}^{+\Delta})]}} - \sum_{\mathbf{f} \in F} e^{-\frac{|S|\epsilon^2}{2Var[L(\mathbf{f},\mathbf{x}^{|S|},w) - l(\mathbf{f},\mathbf{x}^{|S|})]}} - \sum_{k \in K} \sum_{\mathbf{f} \in F} e^{-\frac{|S|\epsilon_k^2}{2Var[L_k(\mathbf{f},\mathbf{x}^{|S|},w) - l_k(\mathbf{f},\mathbf{x}^{|S|})]}}$$

Following the definition of σ^2 and σ_k^2 and taking the summation over $\mathbf{f} \in F$,

$$P(F^{+\Delta} \subseteq F^{|S|} \subseteq F^{-\Delta}) \geq 1 - 2\sum_{\mathbf{f} \in F} e^{-\frac{|S|\epsilon^2}{2\sigma^2}} - 2\sum_{k \in K} \sum_{\mathbf{f} \in F} e^{-\frac{|S|\epsilon_k^2}{2\sigma_k^2}}$$
$$\geq 1 - 2(|K| + 1)|F| e^{-|S| \min\{\frac{\epsilon^2}{2\sigma^2}, \frac{\epsilon_k^2}{2\sigma_k^2}\ k \in K\}}$$

APPENDIX M

CHAPTER 4: COMPARISON OF SCHEDULED AND ACTUAL SURGICAL CASE START TIMES AND DURATIONS

When we compare the scheduled case turnover start and end times with the *actual* start and end times from the surgical data, we observe a deviation in the case start times and durations. To incorporate these (random) deviations to the surgical schedule scenarios as described in Section 4.6.1, we generate empirical distributions for deviations on (i) first case turnover start time, (ii) (non-first) case turnover start time, and (iii) case duration (with turnover). We generate these three distributions for each of the 13 surgical procedure specialties. The reason is that these deviations are sometimes partially explained by the behavior of surgeons or surgical teams working in specific surgical specialties, for instance, by underestimating a case duration to fit it in a surgical schedule [50]. In Figures 47, 48 and 49, we show an example of these three empirical distributions for the orthopedic and colon-rectal specialties.

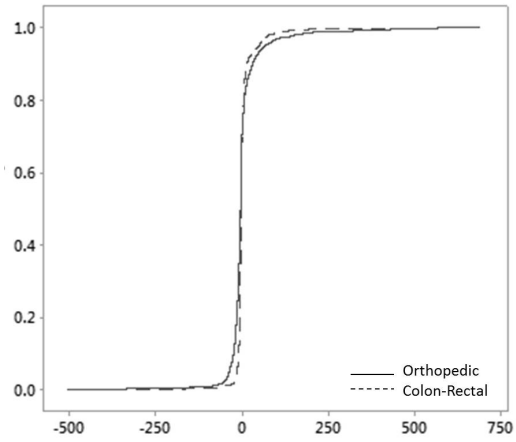


Figure 47: Empirical CDF for the difference between the real and the scheduled first case turnover start time (in minutes) for orthopedic and colon-rectal.

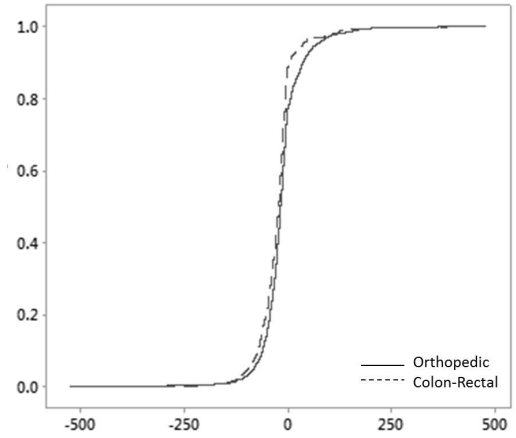


Figure 48: Empirical CDF for the difference between the real and the scheduled (non-first) case turnover start time (in minutes) for orthopedic and colon-rectal.

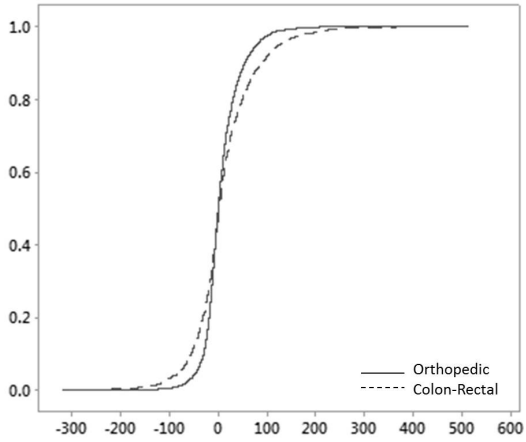


Figure 49: Empirical CDF for the difference between the real and the scheduled case duration with turnover (in minutes) for orthopedic and colon-rectal.

REFERENCES

[1] Agnetis, A., Murgia, G., and Sbrilli, S., "A job shop scheduling problem with human operators in handicraft production," *International Journal of Production Research*, no. ahead-of-print, pp. 1–12, 2013.

[2] Agnihothri, S. and Taylor, P., "Staffing a centralized appointment scheduling department in lourdes hospital," *Interfaces*, vol. 21, pp. 1–11, 1991.

[3] Ahmed, S. and Shapiro, A., "The sample average approximation method for stochastic programs with integer recourse," *Submitted for publication*, 2002.

[4] Alfares, H. K., "An efficient two-phase algorithm for cyclic days-off scheduling," *Computers & Operations Research*, vol. 25, no. 11, pp. 913–923, 1998.

[5] Alfares, H. K., "Survey, categorization, and comparison of recent tour scheduling literature.," *Annals of Operations Research*, vol. 127, no. 1-4, pp. 145–175, 2004.

[6] Alfares, H. K. and Bailey, J. E., "Integrated project task and manpower scheduling," *IIE Transactions*, vol. 29, no. 9, pp. 711–717, 1997.

[7] Andersson, E., Housos, E., Kohl, N., and Wedelin, D., "Crew pairing optimization," in *Operations Research in the Airline Industry*, pp. 228–258, Springer, 1998.

[8] Anitescu, M. and Birge, J., "Convergence of stochastic average approximation for stochastic optimization problems with mixed expectation and per-scenario constraints," *Submitted for publication*, 2009.

[9] Artigues, C., Gendreau, M., Rousseau, L.-M., and Vergnaud, A., "Solving an integrated employee timetabling and job-shop scheduling problem via hybrid branch-and-bound," *Computers & Operations Research*, vol. 36, no. 8, pp. 2330–2340, 2009.

[10] Bachmann, G. A., Trattler, B., Ko, T., and Tweddel, G., "Operational improvement of gynecologic laparoscopic operating room services: an internal review," *Obstetrics & Gynecology*, vol. 92, no. 1, pp. 142–144, 1998.

[11] Baker, K., "Scheduling a full-time workforce to meet cyclic staffing requirements," *Management Science*, vol. 20, pp. 1561–1568, 1974.

[12] Baker, K., "Workforce scheduling with cyclic demands and day-off," *Management Science*, vol. 24, pp. 161–167, 1977.

[13] Balakrishnan, N., Sridharan, V., and Patterson, J. W., "Rationing capacity between two product classes," *Decision Sciences*, vol. 27, no. 2, pp. 185–214, 1996.

[14] Bank of New York Mellon Treasury Services, "How fit is today's healthcare provider's financial back office?: A case for automated EOB processing," 2007.

[15] Bard, J. F., "Staff scheduling in high volume service facilities with downgrading," *IIE Transactions*, vol. 36, no. 10, pp. 985–997, 2004.

[16] Bard, J. F., "Nurse scheduling models," *Wiley Encyclopedia of Operations Research and Management Science*, pp. 3617–3627, 2010.

[17] Bard, J. F., Binici, C., and desilva, A. H., "Staff scheduling at the united states postal service," *Computers & Operations Research*, vol. 30, no. 5, pp. 745–771, 2003.

[18] Bard, J. F., Morton, D. P., and Wang, Y. M., "Workforce planning at USPS mail processing and distribution centers using stochastic optimization," *Annals of Operations Research*, vol. 155, pp. 51–78, 2007.

[19] Bard, J. F. and Purnomo, H. W., "A column generation-based approach to solve the preference scheduling problem for nurses with downgrading," *Socio-Economic Planning Sciences*, vol. 39, no. 3, pp. 193 – 213, 2005.

[20] Bard, J. F. and Wan, L., "Weekly scheduling in the service industry: an application to mail processing and distribution centers," *IIE Transactions*, vol. 37, no. 5, pp. 379–396, 2005.

[21] Beaumont, N., "Scheduling staff using mixed integer programming," *European Journal of Operational Research*, vol. 98, no. 3, pp. 473–484, 1997.

[22] Bechtold, S. E., Brusco, M. J., and Showalter, M. J., "A comparative evaluation of labor tour scheduling methods," *Decision Sciences*, vol. 22, no. 4, pp. 683–699, 1991.

[23] Begur, S. V., Miller, D. M., and Weaver, J. R., "An integrated spatial DSS for scheduling and routing home-health-care nurses," *Interfaces*, vol. 27, no. 4, pp. 35–48, 1997.

[24] Berman, O., Larson, R. C., and Pinker, E., "Scheduling workforce and workflow in a high volume factory," *Management Science*, vol. 43, no. 2, pp. 158–172, 1997.

[25] Branda, M., "Sample approximation technique for mixed-integer stochastic programming problems with expected value constraints," *Optimization Letters*, vol. 8, no. 3, pp. 861–875, 2014.

[26] Brumelle, S. and Granot, D., "The repair kit problem revisited," *Operations Research*, vol. 41, no. 5, pp. 994–1006, 1993.

[27] Brunner, J. O. and Edenharter, G. M., "Long term staff scheduling of physicians with different experience levels in hospitals using column generation," *Health Care Management Science*, vol. 14, no. 2, pp. 189–202, 2011.

[28] Brusco, M. and Jacobs, L., "Starting-time decisions in labor tour scheduling: An experimental analysis and case study," *European Journal of Operational Research*, vol. 131, pp. 459–475, 2001.

[29] Burke, E. K., de Causmaecker, P., Berghe, G. v., and van Landeghem, H., "The state of the art of nurse rostering," *Journal of Scheduling*, vol. 7, no. 6, pp. 441–499, 2004.

[30] Burns, R. and Carter, M., "Work force size and single shift schedules with variable demands," *Management Science*, vol. 31, no. 5, pp. 599–607, 1985.

[31] Camm, J. D., Magazine, M. J., Polak, G. G., and Zaric, G. S., "Scheduling parallel assembly workstations to minimize a shared pool of labor," *IIE Transactions*, vol. 40, no. 8, pp. 749–758, 2008.

[32] Cardoen, B., Demeulemeester, E., and Belien, J., "Operating room planning and scheduling: A literature review," *European Journal of Operational Research*, vol. 201, pp. 921–932, 2010.

[33] Centers for Disease Control and Prevention (CDC), "Guideline for disinfection and sterilization in healthcare facilities, tables," 2008.

[34] Ceran, Y., Dawande, M., Liu, D., and Mookerjee, V., "Optimal software reuse in incremental software development: A transfer pricing approach," *Management Science*, vol. 60, no. 3, pp. 541–559, 2013.

[35] Cezik, T., Gunluk, O., and Luss., H., "An integer programming model for the weekly tour scheduling problem," *Naval Research Logistics*, vol. 48, 2001.

[36] Chen, J. and Askin, R. G., "Project selection, scheduling and resource allocation with time dependent returns," *European Journal of Operational Research*, vol. 193, no. 1, pp. 23–34, 2009.

[37] Chung, J. and White, K., "Cross-trained versus specialized agents in an inbound call centre: a simulation-based methodology for trade-off analysis," *Journal of Simulation*, vol. 2, no. 3, pp. 162–169, 2008.

[38] Cochran, J., Chu, D., and Chu., M., "Optimal staffing for cyclically scheduled processes," *International Journal of Production Research*, vol. 35, pp. 3393–3403, 2001.

[39] Cordone, R., Hosteins, P., Righini, G., Ravizza, P., and Piselli, A., "Optimal selection of contracts and work shifts in multi-skill call centers," *EURO Journal on Computational Optimization*, vol. 2, pp. 247–277, 2014.

[40] Cordone, R., Piselli, A., Ravizza, P., and Righini, G., "Optimization of multi-skill call centers contracts and work-shifts," *Service Science*, vol. 3, no. 1, pp. 67–81, 2011.

[41] Crockett, G. B. and Leamon, P. H., "Skills-based scheduling for telephone call centers," Mar. 28 2000. US Patent 6,044,355.

[42] Dai, L., Chen, C., and Birge, J., "Convergence properties of two-stage stochastic programming," *Journal of Optimization Theory and Applications*, vol. 106, no. 3, pp. 489–509, 2000.

[43] Daniels, R. L. and Mazzola, J. B., "Flow shop scheduling with resource flexibility," *Operations Research*, vol. 42, no. 3, pp. 504–522, 1994.

[44] Dexter, F., "Operating room staffing and allocation (http://www.franklindexter.net/ lectures/orstaffingtalk.pdf)," Sept. 2013.

[45] Dexter, F. and Epstein, R. H., "Optimizing second shift or staffing," *AORN Journal*, vol. 77, no. 4, pp. 825–830, 2003.

[46] Dexter, F. and Epstein, R. H., "Holiday and weekend operating room on-call staffing requirements," *Anesthesia & Analgesia*, vol. 103, no. 6, pp. 1494–1498, 2006.

[47] Dexter, F., Epstein, R. H., Marcon, E., and Ledolter, J., "Estimating the incidence of prolonged turnover times and delays by time of day," *Anesthesiology*, vol. 102, no. 6, pp. 1242–1248, 2005.

[48] Dexter, F., Epstein, R. H., Traub, R. D., and Xiao, Y., "Making management decisions on the day of surgery based on operating room efficiency and patient waiting times," *Anesthesiology*, vol. 101, pp. 1444–1453, 2004.

[49] Dexter, F. and Macario, A., "Applications of information systems to operating room scheduling," *Anesthesiology*, vol. 85, no. 6, pp. 1232–1234, 1996.

[50] Dexter, F., Macario, A., Epstein, R. H., and Ledolter, J., "Validity and usefulness of a method to monitor surgical services average bias in scheduled case durations," *Canadian Journal of Anesthesia*, vol. 52, no. 9, pp. 935–939, 2005.

[51] Dexter, F., Macario, A., and ONeill, L., "Scheduling surgical cases into overflow block timecomputer simulation of the effects of scheduling strategies on operating room labor costs," *Anesthesia & Analgesia*, vol. 90, no. 4, pp. 980–988, 2000.

[52] Dexter, F., Macario, A., Qian, F., and Traub, R. D., "Forecasting surgical groups total hours of elective cases for allocation of block time: application of time series analysis to operating room management," *Anesthesiology*, vol. 91, no. 5, p. 1501, 1999.

[53] Dexter, F. and O'Neill, L., "Weekend operating room on call staffing requirements," *AORN Journal*, vol. 74, no. 5, pp. 664–671, 2001.

[54] Dexter, F. and Traub, R. D., "The lack of systematic month-to-month variation over one-year periods in ambulatory surgery caseloadapplication to anesthesia staffing," *Anesthesia & Analgesia*, vol. 91, no. 6, pp. 1426–1430, 2000.

[55] Di Gaspero, L., Gartner, J., Kortsarz, G., Musliu, N., Schaerf, A., and Slany, W., "The minimum shift design problem," *Annals of Operations Research*, vol. 155, no. 1, pp. 79–105, 2007.

[56] Easton, F. F., "Cross-training performance in flexible labor scheduling environments," *IIE Transactions*, vol. 43, no. 8, pp. 589–603, 2011.

[57] Ernst, A. T., Jiang, H., Krishnamoorthy, M., and Sier, D., "Staff scheduling and rostering: A review of applications, methods and models," *European Journal of Operational Research*, vol. 153, no. 1, pp. 3–27, 2004.

[58] Ernst, A., Jiang, H., Krishnamoorthy, M., Owens, B., and Sier, D., "An annotated bibliography of personnel scheduling and rostering," *Annals of Operations Research*, vol. 127, pp. 21–144, 2004.

[59] Faaland, B. and Schmitt, T., "Cost-based scheduling of workers and equipment in a fabrication and assembly shop," *Operations Research*, vol. 41, no. 2, pp. 253–268, 1993.

[60] Feinberg, E. A. and Yang, F., "Optimality of trunk reservation for an M/M/k/N queue with several customer types and holding costs," *Probability in the Engineering and Informational Sciences*, vol. 25, no. 04, pp. 537–560, 2011.

[61] Firat, M. and Hurkens, C., "An improved MIP-based approach for a multi-skill workforce scheduling problem," *Journal of Scheduling*, vol. 15, no. 3, pp. 363–380, 2012.

[62] Fischetti, M., Lodi, A., Martello, S., and Toth, P., "A polyhedral approach to simplified crew scheduling and vehicle scheduling problems," *Management Science*, vol. 47, no. 6, pp. 833–850, 2001.

[63] Fréville, A., "The multidimensional 0–1 knapsack problem: An overview," *European Journal of Operational Research*, vol. 155, no. 1, pp. 1–21, 2004.

[64] Frost and Sullivan, "Driving operation improvement in hospitals through an advanced instrument management system," 2006.

[65] Fry, M. J., Magazine, M. J., and Rao, U. S., "Firefighter staffing including temporary absences and wastage," *Operations Research*, vol. 54, no. 2, pp. 353–365, 2006.

[66] Güllü, R. and Köksalan, M., "A model for performance evaluation and stock optimization in a kit management problem," *International Journal of Production Economics*, vol. 143, no. 2, pp. 527–535, 2013.

[67] Gupta, D., "Surgical suites' operations management," *Production and Operations Management*, vol. 16, no. 6, pp. 689–700, 2007.

[68] Guyon, O., Lemaire, P., Pinson, E., and Rivreau, D., "Cut generation for an integrated employee timetabling and production scheduling problem," *European Journal of Operational Research*, vol. 201, no. 2, pp. 557–567, 2010.

[69] Guyon, O., Lemaire, P., Pinson, E., and Rivreau, D., "Solving an integrated job-shop problem with human resource constraints," *Annals of Operations Research*, vol. 213, no. 1, pp. 147–171, 2014.

[70] Haase, K., Desaulniers, G., and Desrosiers, J., "Simultaneous vehicle and crew scheduling in urban mass transit systems," *Transportation Science*, vol. 35, no. 3, pp. 286–303, 2001.

[71] Hanssmann, F. and Hess, S. W., "A linear programming approach to production and employment scheduling," *Management Science*, no. 1, pp. 46–51, 1960.

[72] Hartmann, S. and Briskorn, D., "A survey of variants and extensions of the resource-constrained project scheduling problem," *European Journal of Operational Research*, vol. 207, no. 1, pp. 1–14, 2010.

[73] Herbots, J., Herroelen, W., and Leus, R., "Dynamic order acceptance and capacity planning on a single bottleneck resource," *Naval Research Logistics*, vol. 54, no. 8, pp. 874–889, 2007.

[74] Hulshof, P. J., Kortbeek, N., Boucherie, R. J., Hans, E. W., and Bakker, P. J., "Taxonomic classification of planning decisions in health care: a structured review of the state of the art in or/ms," *Health Systems*, vol. 1, no. 2, pp. 129–175, 2012.

[75] Huq, F., Cutright, K., and Martin, C., "Employee scheduling and makespan minimization in a flow shop with multi-processor work stations: A case study," *Omega*, vol. 32, no. 2, pp. 121–129, 2004.

[76] Ingolfsson, A., Haque, A., Umnikov, A., and others, "Accounting for time-varying queueing effects in workforce scheduling," *European Journal of Operational Research*, vol. 139, no. 3, pp. 585–597, 2002.

[77] Jordan, W. C., Inman, R. R., and Blumenfeld, D. E., "Chained cross-training of workers for robust performance," *IIE Transactions*, vol. 36, no. 10, pp. 953–967, 2004.

[78] Kelly, F. P., "Loss networks," *The Annals of Applied Probability*, pp. 319–378, 1991.

[79] Klabjan, D., Johnson, E. L., Nemhauser, G. L., Gelman, E., and Ramaswamy, S., "Airline crew scheduling with time windows and plane-count constraints," *Transportation Science*, vol. 36, no. 3, pp. 337–348, 2002.

[80] Kleywegt, A. J. and Papastavrou, J. D., "The dynamic and stochastic knapsack problem with random sized items," *Operations Research*, vol. 49, no. 1, pp. 26–41, 2001.

[81] Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T., "The sample average approximation method for stochastic discrete optimization," *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 479–502, 2002.

[82] Kohl, N. and Karisch, S. E., "Airline crew rostering: Problem types, modeling, and optimization," *Annals of Operations Research*, vol. 127, no. 1-4, pp. 223–257, 2004.

[83] Kolisch, R. and Meyer, K., "Selection and scheduling of pharmaceutical research projects," in *Perspectives in Modern Project Scheduling*, pp. 321–344, Springer, 2006.

[84] Lee, C.-Y. and Vairaktarakis, G. L., "Workforce planning in mixed model assembly systems," *Operations Research*, vol. 45, no. 4, pp. 553–567, 1997.

[85] Levi, R. and Radovanovic, A., "Provably near-optimal LP-based policies for revenue management in systems with reusable resources," *Operations Research*, vol. 58, no. 2, pp. 503–507, 2010.

[86] Li, L. L. X. and King, B. E., "A healthcare staff decision model considering the effects of staff cross-training," *Health Care Management Science*, vol. 2, no. 1, pp. 53–61, 1999.

[87] Liu, S.-S. and Wang, C.-J., "Optimizing project selection and scheduling problems with time-dependent resource constraints," *Automation in Construction*, vol. 20, no. 8, pp. 1110–1119, 2011.

[88] Louth, G., Mitzenmacher, M., and Kelly, F., "Computational complexity of loss networks," *Theoretical Computer Science*, vol. 125, no. 1, pp. 45–59, 1994.

[89] Maenhout, B. and Vanhoucke, M., "An integrated nurse staffing and scheduling analysis for longer-term nursing staff allocation problems," *Omega*, vol. 41, no. 2, pp. 485–499, 2013.

[90] Masursky, D., Dexter, F., OLeary, C. E., Applegeet, C., and Nussmeier, N. A., "Long-term forecasting of anesthesia workload in operating rooms from changes in a hospitals local population can be inaccurate," *Anesthesia & Analgesia*, vol. 106, no. 4, pp. 1223–1231, 2008.

[91] McIntosh, C., Dexter, F., and Epstein, R. H., "The impact of service-specific staffing, case scheduling, turnovers, and first-case starts on anesthesia group and operating room productivity: a tutorial using data from an australian hospital," *Anesthesia & Analgesia*, vol. 103, no. 6, pp. 1499–1516, 2006.

[92] Mercier, A. and Soumis, F., "An integrated aircraft routing, crew scheduling and flight retiming model," *Computers & Operations Research*, vol. 34, no. 8, pp. 2251–2265, 2007.

[93] MetricNet, "The seven most important performance indicators for service desks," 2009.

[94] Mirus Capital Advisors, "Healthcare providers expected to rely on outsourcing," 2001.

[95] Moore, I. C., Strum, D. P., Vargas, L. G., and Thomson, D. J., "Observations on surgical demand time series: detection and resolution of holiday variance," *Anesthesiology*, vol. 109, no. 3, pp. 408–416, 2008.

[96] Morris, J. G. and Showalter, M. J., "Simple approaches to shift, days-off and tour scheduling problems," *Management Science*, vol. 29, no. 8, pp. 942–950, 1983.

[97] Munro, D., "Annual u.s. healthcare spending hits $3.8 trillion," *Forbes*, February 2014.

[98] Neumann, K. and Zimmermann, J., "Resource levelling for projects with schedule-dependent time windows," *European Journal of Operational Research*, vol. 117, no. 3, pp. 591–605, 1999.

[99] Ni, J., Tsang, D. H., Tatikonda, S., and Bensaou, B., "Optimal and structured call admission control policies for resource-sharing systems," *IEEE Transactions on Communications*, vol. 55, no. 1, pp. 158–170, 2007.

[100] O'Donnell, J., "Health care spending growth is slow but rising," *USA Today*, September 2014.

[101] Peng, F. and Ouyang, Y., "Track maintenance production team scheduling in railroad networks," *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1474–1488, 2012.

[102] Presidents Council of Advisors on Science and Technology (PCAST), "Better health care and lower costs: Accelerating improvement through systems engineering," 2008.

[103] Puhalskii, A. A. and Reiman, M. I., "A critically loaded multirate link with trunk reservation," *Queueing Systems*, vol. 28, no. 1-3, pp. 157–190, 1998.

[104] Rutala, W. A. and Weber, D. J., "Disinfection and sterilization in health care facilities: what clinicians need to know," *Clinical Infectious Diseases*, vol. 39, no. 5, pp. 702–709, 2004.

[105] Sayin, S. and Karabati, S., "Assigning cross-trained workers to departments: A two-stage optimization model to maximize utility and skill improvement," *European Journal of Operational Research*, vol. 176, no. 3, pp. 1643–1658, 2007.

[106] Senju, S. and Toyoda, Y., "An approach to linear programming with 0-1 variables," *Management Science*, pp. B196–B207, 1968.

[107] Shapiro, A. and Homem-de Mello, T., "On the rate of convergence of optimal solutions of monte carlo approximations of stochastic programs," *SIAM Journal on Optimization*, vol. 11, no. 1, pp. 70–86, 2000.

[108] Sinreich, D. and Jabali, O., "Staggered work shifts: a way to downsize and restructure an emergency department workforce yet maintain current operational performance," *Health Care Management Science*, vol. 10, no. 3, pp. 293–308, 2007.

[109] Slotnick, S. A., "Order acceptance and scheduling: a taxonomy and review," *European Journal of Operational Research*, vol. 212, no. 1, pp. 1–11, 2011.

[110] Society for Human Resource Management, "Salaries as a percentage of operating expense," Nov. 2008.

[111] Taylor, S. J., Brailsford, S., Chick, S. E., L'Ecuyer, P., Macal, C. M., and Nelson, B. L., "Modeling and simulation grand challenges: An OR/MS perspective," in *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, pp. 1269–1282, IEEE Press, 2013.

[112] Teunter, R. H., "The multiple-job repair kit problem," *European Journal of Operational Research*, vol. 175, no. 2, pp. 1103–1116, 2006.

[113] Tien, J. M. and Kamiyama, A., "On manpower scheduling algorithms," *SIAM Review*, vol. 24, no. 3, pp. 275–287, 1982.

[114] van den Bergh, J., Beliën, J., de Bruecker, P., Demeulemeester, E., and de Boeck, L., "Personnel scheduling: A literature review," *European Journal of Operational Research*, vol. 226, no. 3, pp. 367–385, 2013.

[115] van Oostrum, J. M., Bredenhoff, E., and Hans, E. W., "Suitability and managerial implications of a master surgical scheduling approach," *Annals of Operations Research*, vol. 178, pp. 91–104, 2010.

[116] van Oostrum, J. M., van Houdenhoven, M., Vrielink, M. M., Klein, J., Hans, E. W., Klimek, M., Wullink, G., Steyerberg, E. W., and Kazemier, G., "A simulation model for determining the optimal size of emergency teams on call in the operating room at night," *Anesthesia & Analgesia*, vol. 107, no. 5, pp. 1655–1662, 2008.

[117] Viapiano, J. and Ward, D. S., "Operating room utilization: the need for data," *International Anesthesiology Clinics*, vol. 38, no. 4, pp. 127–140, 2000.

[118] Vliegen, I. and van Houtum, G., "Approximate evaluation of order fill rates for an inventory system of service tools," *International Journal of Production Economics*, vol. 118, no. 1, pp. 339–351, 2009.

[119] Vohra, R. V., "The cost of consecutivity in the (5,7) cyclic staffing problem," *IIE Transactions*, vol. 19, no. 3, pp. 296–299, 1987.

[120] Walters, E., "Two sample kolmogorov-smirnov test of the underyling distributions code," 2009.

[121] Wang, J., Quan, S., Li, J., and Hollis, A. M., "Modeling and analysis of work flow and staffing level in a computed tomography division of university of wisconsin medical foundation," *Health Care Management Science*, vol. 15, no. 2, pp. 108–120, 2012.

[122] Wang, W. and Ahmed, S., "Sample average approximation of expected value constrained stochastic programs," *Operations Research Letters*, vol. 36, no. 5, pp. 515–519, 2008.

[123] Weingartner, H. M. and Ness, D. N., "Methods for the solution of the multidimensional 0/1 knapsack problem," *Operations Research*, vol. 15, no. 1, pp. 83–103, 1967.

VITA

Monica Villarreal was born in Monterrey, Mexico, where she received her B.S. in Chemical Engineering from Tec de Monterrey in December 2003. She joined the logistics unit of FEMSA, the biggest beverage company in Latin America, before coming to Georgia Tech and completing her M.S. in Industrial Engineering in August 2006. Before returning to Georgia Tech for her Ph.D. in the fall of 2008, she worked as a supply chain consultant in Latin America. Her research interests include humanitarian and healthcare applications of operations research and management science. As a Ph.D. student, she has had the opportunity to work with regional hospitals such as Childrens Healthcare of Atlanta and Emory University Hospital in Atlanta, Georgia, and University Hospital in Augusta, Georgia on process improvement and delays/waste reduction, staff planning, etc. She has also worked with the Health and Humanitarian Logistics Center at Georgia Tech, in projects related to operations research applications in humanitarian logistics, the role of the private sector in disaster response, debris collection planning, and the logistics behind cholera prevention and response.